NeuroImage 170 (2018) 482–494

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement

Roberto Souza [a,c,d,e,*], Oeslle Lucena [a], Julia Garrafa [b], David Gobbi [c,d], Marina Saluzzi [c,d], Simone Appenzeller [b], Letícia Rittner [a], Richard Frayne [c,d,e], Roberto Lotufo [a]

[a] Medical Imaging and Computing Laboratory, Department of Computer Engineering and Industrial Automation, University of Campinas, Campinas, São Paulo, Brazil
[b] Division of Rheumatology, Faculty of Medical Science, University of Campinas, Campinas, São Paulo, Brazil
[c] Departments of Radiology and Clinical Neurosciences, Hotchkiss Brain Institute, University of Calgary, Calgary, Alberta, Canada
[d] Calgary Image Processing and Analysis Centre, Foothills Medical Centre, Alberta Health Services, Calgary, Alberta, Canada
[e] Seaman Family Magnetic Resonance Research Centre, Foothills Medical Centre, Alberta Health Services, Calgary, Alberta, Canada

ARTICLE INFO

Article history:
Received 1 November 2016
Received in revised form
4 August 2017
Accepted 5 August 2017
Available online 12 August 2017

Keywords:
Public database
Skull stripping
Brain extraction
Brain segmentation
Brain MR image analysis
MP-RAGE

ABSTRACT

This paper presents an open, multi-vendor, multi-field strength magnetic resonance (MR) T1-weighted volumetric brain imaging dataset, named *Calgary-Campinas-359* (*CC-359*). The dataset is composed of images of older healthy adults (29–80 years) acquired on scanners from three vendors (Siemens, Philips and General Electric) at both 1.5 T and 3 T. *CC-359* is comprised of 359 datasets, approximately 60 subjects per vendor and magnetic field strength. The dataset is approximately age and gender balanced, subject to the constraints of the available images. It provides consensus brain extraction masks for all volumes generated using supervised classification. Manual segmentation results for twelve randomly selected subjects performed by an expert are also provided. The *CC-359* dataset allows investigation of 1) the influences of both vendor and magnetic field strength on quantitative analysis of brain MR; 2) parameter optimization for automatic segmentation methods; and potentially 3) machine learning classifiers with big data, specifically those based on deep learning methods, as these approaches require a large amount of data. To illustrate the utility of this dataset, we compared to the results of a supervised classifier, the results of eight publicly available skull stripping methods and one publicly available consensus algorithm. A linear mixed effects model analysis indicated that vendor ($p - value < 0.001$) and magnetic field strength ($p - value < 0.001$) have statistically significant impacts on skull stripping results.

1. Introduction

Magnetic resonance (MR) imaging is an important tool in the diagnosis and follow-up care of patients with brain disease and disorders. More specifically, quantitative brain image analysis is playing an increasingly important role in the development and execution of clinical and research studies. Skull stripping, also known as brain extraction or brain segmentation, is the process of segmenting brain from non-brain tissue. In MR images, skull stripping is an initial step for many quantitative image analysis applications, such as multi-modal registration, cortical flattening procedures, and brain atrophy estimation (Smith, 2002). Brain extraction is an active research field (Avants et al., 2011; Iglesias et al., 2011; Beare et al.,; Eskildsen et al., 2012; Kleesiek et al., 2016). To date, there are four main classes of methods proposed for performing skull stripping: 1) manual segmentation, 2) intensity-based models associated with morphology, 3) surface model-based, and 4) hybrid methods.

Manual segmentations are frequently considered to be the "gold-standard" for skull stripping and are often used to validate other automatic and semi-automatic methods. This method, however, is labor intensive, and therefore impractical in large datasets. Intensity-based methods, such as the ones that use the watershed transform (Beare et al.,; Hahn and Peitgen, 2000), require less computational time compared to other methods and are able to include the brain stem, spinal cord, and much of the brain gyral surface in their segmentation results. Unfortunately, intensity-based methods often produce over-segmented results, *i.e.* results where the structure of interest is split in two or more regions in the final segmentation mask. Model-based

* Corresponding author. Medical Imaging and Computing Laboratory, Department of Computer Engineering and Industrial Automation, University of Campinas, Campinas, São Paulo, Brazil.
E-mail addresses: roberto.medeirosdeso@ucalgary.ca (R. Souza)
URL: http://miclab.fee.unicamp.br

https://doi.org/10.1016/j.neuroimage.2017.08.021
1053-8119/

methods (Smith, 2002) use a balloon-like template, which is fit to the brain surface using gradient information and smoothing forces. Some model-based methods require registration to an atlas as a pre-processing step and therefore have longer processing times than intensity-based methods. In addition, due to smoothness constraints, they are not typically able to include the spinal cord and brain stem in the segmentation result. As they generate a smoothed surface, they are also not able to properly segment the brain gyral surface. Hybrid methods attempt to combine the best features of intensity-based and model-based methods, require longer processing times, but achieve improved segmentation results (Ségonne, et al.).

Validating automatic and semi-automatic brain extraction methods is a difficult task and often requires comparison against manually segmented data and only a few public datasets include manual segmentation results. Simulated T1-weighted MR images can also be used for validating automatic methods (Lee et al., 2003). For validation of skull stripping methods, the commonly used datasets are those from the Open Access Series of Imaging Studies (OASIS) (Marcus et al., 2007) and the LONI Probabilistic Brain Atlas (LPBA40) (Shattuck et al., 2008). OASIS includes 77 subjects, of which 20 are classified as being cognitively impaired. For each subject in OASIS, three to four T1-weighted 3D MP-RAGE scans were acquired and co-registered. All of the images were collected on a 1.5 T Siemens scanner with a slice thickness of 1.25 mm. The LPBA40 dataset includes 40 coronal 3D T1-weighted spoiled gradient echo MR scans acquired on a 1.5 T General Electric (GE) scanner. The slice thickness of the images was 1.5 mm. These publicly available datasets are relatively small in size and typically do not allow for analysis of other important acquisition parameters, such as scanner vendor and/or magnetic field strength. Nevertheless, in the absence of manually segmented data, while it is not possible to fully characterize segmentation performance, it is possible to detect outliers, and evaluate the overall consistency and similarity between different techniques (Bouix et al., 2007).

Presented in this paper is a multi-vendor, multi-field strength database. The utility of this database is demonstrated by evaluating the agreement of a series of eight publicly available skull stripping methods. In addition, consensus segmentation masks were generated for each subject using the Simultaneous Truth and Performance Level Estimation (STAPLE) method (Warfield et al., 2004). Manual segmentation was performed on a subset of twelve subjects and with a supervised classifier, used to develop what we will call"silver standard" (SS) brain masks across all subjects in order to assess agreement in our study.

Our results show that scanner vendor and magnetic field strength significantly influence the skull stripping results. To the best of our knowledge, this effort is the first work that analyzes the influences of both scanner vendor and magnetic field strength on skull stripping. Previous work has assessed skull stripping performance in data acquired on different scanners at different institutions (Boesen et al., 2004), but used private data, therefore preventing full assessment of the robustness of these studies with respect to vendor and magnetic field strength. Our dataset is publicly accessible, and can be used to optimize skull stripping parameters depending on scanner vendor and magnetic field strength. Also, the dataset can be used to increase the amount of data necessary to train approaches based on deep learning (Kleesiek et al., 2016).

2. Materials and methods

2.1. Public dataset - the Calgary-Campinas-359

The public dataset we have developed consists of T1 volumes acquired in 359 subjects on scanners from three different vendors (GE, Philips, and Siemens) and at two magnetic field strengths (1.5 T and 3 T). Data was obtained using T1-weighted 3D imaging sequences (3D MP-RAGE (Philips, Siemens), and a comparable T1-weighted spoiled gradient echo sequence (GE)) designed to produce high-quality anatomical data with 1 mm^3 voxels. Older adult subjects were scanned between 2009 and 2016.

Smaller, private datasets in Campinas, São Paulo, Brazil and Calgary, Alberta, Canada were used to randomly select the 359 subjects, except for the Philips 1.5 T data where only 59 subjects were available. Age and gender for all subjects were known, however information about subject ethnicity was not available. The *Calgary-Campinas-359* (*CC-359*) dataset, including the original Nifti files (Cox et al., 2004), the consensus masks generated for all subjects using both the STAPLE algorithm and a supervised classification procedure and the manual segmentations of twelve subjects, are available for download (http://miclab.fee.unicamp.br/tools). Detailed information about the acquisition parameters, such as echo time, repetition time, *etc*., can be provided upon request.

2.2. Manual segmentation

Twelve subjects, two for each vendor-magnetic field strength combination, were randomly selected and then manually segmented using ITK-snap (Yushkevich et al., 2006) available at the Insight Segmentation and Registration Toolkit (ITK) repository. The segmentations were performed in three stages. First, one individual reviewed the axial slices from inferior to superior and voxels corresponding to brain were coarsely marked. In the second stage, a second individual reviewed the reformatted sagittal slices to refine the coarse segmentation. The final stage consisted of a third individual reviewing and providing fine corrections to the segmentation. No automated segmentation method was used to seed the manual segmentation. Each volume took roughly five hours to segment (first two stages). The final stage review required on average twenty minutes per volume.

2.3. Skull stripping techniques

In this study, we compared eight commonly used skull stripping techniques that have publicly available source code. A ninth, consensus building, technique was also assessed. We recognize that there are a number of more recent techniques (*c.f.*, (Khastavaneh and Ebrahimpour-Komleh, 2015; Kleesiek et al., 2016; Roy and Maji, 2015)) but these methods did not meet the source code availability criterion. The nine techniques we compared in this study, in alphabetical order, were:

- Advanced Normalization Tools (ANTs) (Avants et al., 2011): uses non-linear registration to register a brain atlas to the subject space and mask out the background. The default non-linear registration parameters were used.
- Brain Extraction based on non-local Segmentation Technique (BEaST) (Eskildsen et al., 2012): based on non-local segmentation embedded in a multi-resolution framework. We used the default parameters.
- Brain Extraction Tool (BET) (Smith, 2002) from FSL (Jenkinson et al., 2012) software: uses a deformable model that evolves to fit the brain surface by the application of a set of locally adaptive model forces. Two main parameters are: fractional intensity threshold and vertical gradient in fractional intensity threshold. We used the same parameters used in (Iglesias et al., 2011).
- Brain Surface Extractor (BSE) (Shattuck et al., 2001) from BrainSuite (Shattuck and Leahy, 2000) software: combines anisotropic diffusion filtering, edge detection, and mathematical morphology. It has many parameters that require fine tuning. We used the default parameters.
- Hybrid Watershed Approach (HWA) (Ségonne, et al.,) from Freesurfer (Dale et al., 1999) software: combines a model-based method and the watershed transform to segment the brain. The

pre-flooding height used by the watershed transform can be set, but it is usually robust when the default values are employed.

- Marker Based Watershed Scalper (MBWSS) (Beare et al.,): uses the watershed transform from markers and aggressive filtering with large kernels. Many parameter adjustments are possible; however, we followed the authors' recommendation to use the default values.
- Optimized Brain Extraction (OPTIBET) (Lutkenhoff et al., 2014): combines non-linear registration with the previously described BET algorithm (Smith, 2002). The default non-linear registration parameters were used.
- Robust Brain Extraction (ROBEX) (Iglesias et al., 2011): combines a discriminative and a generative model to extract the brain. The discriminative model is a random forest classifier (Breiman, 2001) and the samples used to train it come from images acquired on a Bruker 4 T system (Iglesias et al., 2011). No parameters to set.
- Simultaneous Truth and Performance Level Estimation (STAPLE) (Warfield et al., 2004) is a consensus forming algorithm that uses the results of two or more brain extraction techniques and an expectation-maximization algorithm to estimate the true segmentation. The algorithm is available at the ITK repository (Ibáñez et al., 2003). For this study, the STAPLE algorithm found a consensus mask using as input the segmentations of the eight previously described skull-stripping techniques.

2.4. Supervised classification consensus

Consensus methods combine different segmentations with the objective of obtaining more accurate and robust results (Warfield et al., 2004; Asman and Landman, 2011; Rehm et al., 2004; Rex et al., 2004). Rex et al. (2004), for example, compared the results of their consensus method when combining automatic methods. In this study, they obtained a higher agreement rate than the individual segmentations made by two experts. STAPLE, one of the assessed methods, can form a consensus between masks but assumes the input data to be uncorrelated (Warfield et al., 2004).

In this work, a supervised classifier approach was used to generate a second set of consensus masks between the eight described skull-stripping methods. The subset of manually segmented data was used to train a supervised classifier, which in our case was a logistic regression classifier (Collins et al., 2002). Our feature vector is composed of eight boolean variables (the outputs of each segmentation method). This allows for 256 different combinations of features, which is a relatively small search space. During our tests, we tried other classifiers, such as decision tree (Breiman et al., 1984), random forest (Breiman, 2001) and support vector machine (Steinwart and Christmann, 2008), but all classifiers achieved similar results. We opted to use logistic regression because it is both simple and fast. We did not use all the image voxels to train the classifier, we subtracted the erosion (Soille, 2004) of the binary manual segmentation mask from the original image to extract the border. Then, we perform a dilation (Soille, 2004) with a given radius (experimentally set) of the extracted border. The resulting non-zero voxels were used as training samples for the classifier. This procedure allowed the classifier to learn from the most difficult samples, which are the ones close to the border. The training samples extraction and the construction of the feature vectors are summarized in Fig. 1. Although the classifier is trained only on the border voxels, the final classification was done on the entire image volume.

Our supervised classifier was validated using a 2-fold cross-validation procedure (6 manually segmented images per fold). Then, the 12 subjects with manual segmentation were used to train the final classifier. The resulting supervised classifier-obtained consensus was considered as a “silver standard”, and was used to evaluate the agreement between the skull-stripping methods. Supervised classification approaches have the advantage that they can learn and then correct for methodological and other correlations in the inputs. Similar approaches have been used in other studies to improve skull-stripping (Rehm et al., 2004; Rex et al., 2004; Wang and Yushkevich, 2013). Wang and Yushkevich (2013) proposed a corrective learning approach that uses an AdaBoost classifier (Friedman, 2001) to fix possible systematic segmentation errors that may occur due to limitations of the segmentation model or due to suboptimal solutions obtained by the segmentation optimization algorithm. Although their corrective learning is not a consensus, you can pass segmentation masks as additional features, therefore making the result a consensus. We considered using their approach, but initial results of our method were superior when compared to their results.

2.5. Evaluation criteria and statistical analysis

2.5.1. Evaluation metrics

The metrics used to evaluate the segmentation results were: Dice coefficient, sensitivity, specificity, Hausdorff distance, and mean symmetric surface-to-surface distance. The first three metrics are overlap metrics and the last two, border distance metrics. All metrics are commonly used to analyze skull stripping segmentation performance (Iglesias et al., 2011; Eskildsen et al., 2012; Beare et al.,). Suppose that G is the ground truth image and S is the segmentation we want to assess, the metrics are given by the following equations:

- Dice coefficient:

$$Dice(G, S) = \frac{2|S \cap G|}{|S| + |G|}$$

- Sensitivity:

$$Sensitivity(G, S) = \frac{|G \cap S|}{|G|}$$

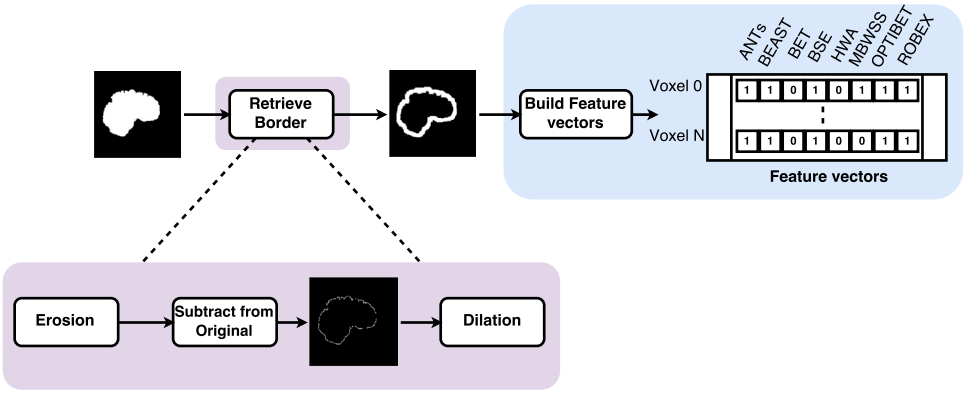


Fig. 1. Summary of our training samples extraction and feature vectors construction. The purple box illustrates the procedure for extracting training samples near the border. The blue box illustrates how the feature vectors are built.

Table 1
Dataset description. Columns from left to right: scanner vendor, magnetic field strength, average age (mean ± standard deviation), gender (number of male/number of female subjects), number of subjects included and number of manual segmentations available.

Vendor	Field	Age	Gender	#subjects	#manual
Siemens	1.5 T	53.9 ± 7.3	30M/30F	60	2
	3.0 T	56.6 ± 6.9	30M/30F	60	2
Philips	1.5 T	52.8 ± 9.6	26M/33F	59	2
	3.0 T	50.0 ± 9.3	30M/30F	60	2
GE	1.5 T	53.9 ± 5.8	30M/30F	60	2
	3.0 T	53.6 ± 5.7	30M/30F	60	2
All	1.5 T and 3 T	53.5 ± 7.8	176M/183F	359	12

- Specificity:

$$Specificity(G, S) = \frac{|G^c \cap S^c|}{|G^c|}$$

- Haussdorff distance:

$$d_H(S, G) = \max\left\{ \sup_{s\in S} \inf_{g\in G} d(s, g), \sup_{g\in G} \inf_{s\in S} d(s, g) \right\}$$

- Symmetric surface-to-surface mean distance:

$$d_S(S, G) = \frac{\sum_{s\in S} \min_{g\in G} d(s, g) + \sum_{g\in G} \min_{s\in S} d(g, s)}{|S| + |G|}$$

The Dice coefficient can be viewed as a compromise between sensitivity and specificity and is probably the most widely used metric to assess segmentation. Sensitivity measures how much brain tissue is left out of the segmentation. Specificity measures how much non-brain tissue is included in the segmentation. The Haussdorff distance is an indicative of outliers and the symmetric surface-to-surface mean distance is similar to the Dice coefficient, but easier to interpret.

2.5.2. Statistical analysis

Variations in the age and gender distribution between the three scanner vendors and two field strengths were assessed using analysis of variance (ANOVA) and chi-squared tests, respectively. Wilcoxon signed-rank tests with Bonferroni correction were used to assess differences in the evaluation metrics. Wilcoxon signed-rank test is a non-parametric statistical hypothesis test that does not assume normal distribution (Haynes, 2013). A p-value < 0.05 was used to assess statistical significance.

The influences of magnetic field intensity, scanner vendor and gender were analyzed using a linear mixed effects (LME) model using the *lme4* package (Bates et al., 2015). One of the major advantages of a LME model is that it does not assume independence among observations. Dice coefficient was used as the dependent variable. As fixed effects, we entered magnetic field strength, scanner vendor and gender. As random effects, we had intercepts for the different skull stripping methods and age. For the LME model, any residual value greater than two standard deviations from the mean (i.e., outside the 95% confidence interval) was deemed to be significantly different.

3. Results

3.1. Public dataset - the Calgary-Campinas-359 dataset characteristics

Average age of the subjects in the *CC-359* database was 53.5 ± 7.8 years (mean ± standard deviation) with an age range from 29 to 80 years. General demographic information on the dataset is summarized in Table 1. The database included 183 (50.97%) female subjects (55.5 ± 7.0 years; range: 36–80 years) and 176 (49.03%) male subjects (51.4 ± 8.1 years; range: 29–71 years). A significant difference in age distribution ($p - value < 0.001$, ANOVA) was found. Post-hoc testing with Bonferroni correction demonstrated that only the Philips 3 T and Siemens 3 T group age

Table 2
Overall analysis (Dice coefficient, sensitivity, specificity, Hausdorff distance and mean symmetric distance). The two best scores for each metric are emboldened.

Method	Dice	Sensitivity	Specificity	Hausdorff	Mean distance
ANTS	**97.587 ± 1.014**	96.698 ± 1.383	99.821 ± 0.196	**8.772 ± 4.058**	**0.038 ± 0.038**
BEaST	97.357 ± 1.107	95.561 ± 2.038	99.913 ± 0.141	9.615 ± 8.424	0.041 ± 0.035
BET	93.877 ± 8.859	98.436 ± 5.517	98.567 ± 2.592	17.303 ± 20.280	0.675 ± 3.188
BSE	90.065 ± 15.562	90.199 ± 13.774	98.405 ± 4.220	54.870 ± 31.428	1.894 ± 5.989
HWA	91.283 ± 1.243	**99.995 ± 0.019**	97.715 ± 0.666	15.588 ± 5.952	0.213 ± 0.047
MBWSS	96.906 ± 4.199	94.632 ± 6.467	**99.950 ± 0.123**	25.335 ± 9.349	0.159 ± 0.625
OPTIBET	96.564 ± 0.705	97.409 ± 0.720	99.477 ± 0.269	12.340 ± 9.114	0.054 ± 0.016
ROBEX	95.748 ± 1.042	98.690 ± 0.696	99.120 ± 0.326	10.559 ± 3.399	0.078 ± 0.049
STAPLE	**97.657 ± 1.063**	**99.860** ± 0.151	99.475 ± 0.244	**6.827 ± 2.071**	**0.014 ± 0.010**

Table 3
Summary of Wilcoxon signed-rank tests with Bonferroni correction. The elements above the main diagonal correspond to the comparison against the "silver standard" (SS) and the elements below the main diagonal correspond to the comparison against the manual segmentations. The statistical significance of the five performance metrics (ordered from left to right: Dice coefficient, sensitivity, specificity, Haussdorff distance and symmetric surface-to-surface mean distance) are reported using "**Y**" to denote that a statistical significant difference was found and "N" to denote that it was not. Where "Y" is underlined, it means that in the pairwise comparison, the method in the corresponding column was better than the method in the corresponding row. When, "Y" is not underlined, it means the contrary.

	ANTs	BEaST	BET	BSE	HWA	MBWSS	OPTIBET	ROBEX	STAPLE	SS
ANTs	–	**YY**Y**YY**	**Y**Y**YYY**	**YYYYY**	**Y**Y**YYY**	N**Y**Y**YY**	**Y**Y**YYY**	**Y**Y**YYY**	N**YY**Y**Y**	–
BEaST	NNNNN	–	**Y**Y**YYY**	**YYYYY**	**Y**Y**YYY**	NN**YYY**	**Y**Y**YYY**	**Y**Y**YYY**	**YYY**Y**Y**	–
BET	N**Y**Y**Y**N	N**Y**Y**N**N	–	**YYYYY**	**Y**Y**Y**Y**Y**	**YY**Y**Y**Y	**YY**Y**YY**	N**Y**N**Y**N	**YYYYY**	–
BSE	NNN**YY**	**Y**N**YYY**	**YY**N**YY**	–	**YYY**Y**N**	**YYYYY**	**YYYYY**	**YYYYY**	**YYYYY**	–
HWA	**Y**Y**Y** Y**Y**	**YY**Y**YY**	**YY**Y**N**Y	N**Y**N**Y**N	–	**YY**Y**Y**Y	**YY**Y**YY**	**YY**Y**YY**	**YY**Y**YY**	–
MBWSS	NN**Y**Y**N**	NNN**Y**N	N**YY**Y**N**	NN**YYY**	**Y**Y**Y**Y**Y**	–	**Y**Y**YY**	**Y**Y**YY**	N**YY**Y**Y**	–
OPTIBET	N**Y**Y**NN**	NN**Y**NN	NNNNN	NNN**YY**	**Y**Y**YYY**	N**Y**Y**Y**N	–	**Y**Y**Y**Y**Y**	**YY**N**YY**	–
ROBEX	N**Y**Y**NN**	N**Y**Y**NN**	NNNNN	**YY**N**YY**	**Y**Y**YYY**	N**Y**Y**Y**N	N**Y**NNN	–	**YYYYY**	–
STAPLE	N**Y**Y**N**Y	N**Y**Y**N**Y	**Y**N**YYY**	**YY**N**YY**	**Y**Y**YYY**	N**Y**Y**YY**	**YY**NN**Y**	**Y**N**Y**N**Y**	–	–
SS	**YY**NN**Y**	**YY**N**YY**	**Y**N**YYY**	**YY**N**YY**	**Y**Y**YYY**	**YY**Y**YY**	**Y**N**YYY**	**Y**Y**Y**N**Y**	N**YY**NN	–

Table 4
Analysis by vendor (Dice, sensitivity, specificity, Hausdorff distance, and mean symmetric distance). The best score for each scanner vendor and metric are emboldened.

Method	Vendor	Dice	Sensitivity	Specificity	Hausdorff	Mean distance
ANTs	Philips	97.488 ± 1.161	96.664 ± 1.820	99.767 ± 0.191	8.198 ± 2.293	0.039 ± 0.044
	Siemens	**97.963 ± 0.631**	97.154 ± 0.901	99.883 ± 0.104	9.071 ± 6.207	0.031 ± 0.019
	GE	97.309 ± 1.055	96.276 ± 1.114	99.813 ± 0.248	9.043 ± 2.245	0.046 ± 0.043
BEaST	Philips	96.895 ± 1.598	94.595 ± 3.076	99.914 ± 0.115	8.674 ± 1.728	0.047 ± 0.031
	Siemens	97.502 ± 0.779	95.798 ± 0.724	99.924 ± 0.206	9.787 ± 7.577	0.040 ± 0.051
	GE	97.672 ± 0.440	96.283 ± 1.004	99.900 ± 0.060	10.376 ± 12.265	0.035 ± 0.009
BET	Philips	94.012 ± 11.520	96.670 ± 9.096	98.572 ± 2.858	16.767 ± 22.617	0.990 ± 4.655
	Siemens	94.064 ± 4.487	99.771 ± 0.321	98.821 ± 1.144	17.117 ± 14.289	0.318 ± 1.573
	GE	93.556 ± 9.105	98.852 ± 1.969	98.307 ± 3.250	18.019 ± 22.764	0.720 ± 2.494
BSE	Philips	92.395 ± 9.497	88.972 ± 9.101	99.397 ± 2.306	39.571 ± 25.494	0.698 ± 3.937
	Siemens	94.020 ± 12.337	91.828 ± 12.363	99.650 ± 1.199	54.085 ± 19.349	0.694 ± 4.167
	GE	83.800 ± 20.558	89.786 ± 18.119	96.175 ± 6.250	70.827 ± 38.061	4.279 ± 8.126
HWA	PHILIPS	90.773 ± 1.505	**99.997 ± 0.013**	97.062 ± 0.534	15.372 ± 1.752	0.227 ± 0.054
	Siemens	91.461 ± 1.088	**99.994 ± 0.025**	98.185 ± 0.516	16.057 ± 7.235	0.214 ± 0.044
	GE	91.613 ± 0.892	**99.994 ± 0.019**	97.891 ± 0.322	15.334 ± 7.089	0.199 ± 0.035
MBWSS	Philips	96.654 ± 6.203	94.558 ± 9.373	**99.928 ± 0.073**	29.075 ± 10.392	0.257 ± 0.986
	Siemens	96.392 ± 3.518	93.543 ± 5.544	**99.960 ± 0.197**	23.080 ± 9.593	0.170 ± 0.418
	GE	97.671 ± 1.141	95.795 ± 2.170	**99.961 ± 0.022**	23.881 ± 6.426	0.050 ± 0.091
OPTIBET	Philips	96.427 ± 0.942	97.676 ± 0.783	99.299 ± 0.333	11.453 ± 4.746	0.057 ± 0.023
	Siemens	96.728 ± 0.535	97.399 ± 0.651	99.607 ± 0.168	13.403 ± 11.062	0.053 ± 0.012
	GE	96.537 ± 0.525	97.155 ± 0.621	99.524 ± 0.168	12.157 ± 10.092	0.054 ± 0.011
ROBEX	Philips	95.716 ± 1.132	98.699 ± 0.680	98.912 ± 0.326	10.097 ± 2.616	0.075 ± 0.037
	Siemens	95.601 ± 0.895	99.098 ± 0.426	99.212 ± 0.245	11.029 ± 3.498	0.083 ± 0.040
	GE	95.927 ± 1.059	98.273 ± 0.683	99.235 ± 0.298	10.546 ± 3.886	0.075 ± 0.064
STAPLE	Philips	**97.928 ± 1.159**	99.806 ± 0.185	99.429 ± 0.324	**6.815 ± 2.135**	**0.016 ± 0.016**
	Siemens	97.156 ± 1.065	99.906 ± 0.121	99.473 ± 0.187	**6.424 ± 1.779**	**0.012 ± 0.003**
	GE	**97.889 ± 0.734**	99.867 ± 0.121	**99.521 ± 0.188**	**7.241 ± 2.194**	**0.014 ± 0.004**

Table 5
Analysis by magnetic field strength (Dice, sensitivity, specificity, Hausdorff distance and mean symmetric distance). The best score for each magnetic field strength and metric are highlighted emboldened.

Method	Strength	Dice	Sensitivity	Specificity	Hausdorff	Mean distance
ANTs	1.5 T	97.502 ± 0.959	96.605 ± 1.173	99.803 ± 0.202	8.816 ± 5.251	0.039 ± 0.036
	3 T	**97.672 ± 1.059**	96.791 ± 1.559	99.838 ± 0.189	8.729 ± 2.328	0.037 ± 0.039
BEaST	1.5 T	97.666 ± 0.770	96.241 ± 1.044	99.893 ± 0.172	9.114 ± 6.369	0.035 ± 0.043
	3 T	97.051 ± 1.290	94.886 ± 2.507	99.932 ± 0.0989	10.113 ± 10.034	0.046 ± 0.024
BET	1.5 T	94.814 ± 7.401	98.630 ± 1.672	98.703 ± 2.557	14.729 ± 18.778	0.493 ± 2.059
	3 T	92.946 ± 10.016	98.243 ± 7.607	98.431 ± 2.620	19.862 ± 21.366	0.855 ± 3.999
BSE	1.5 T	93.520 ± 10.375	90.596 ± 11.105	99.654 ± 0.983	39.383 ± 24.435	0.552 ± 3.421
	3 T	86.630 ± 18.770	89.804 ± 15.983	97.162 ± 5.608	70.271 ± 30.008	3.227 ± 7.506
HWA	1.5 T	91.317 ± 0.975	**99.995 ± 0.016**	97.577 ± 0.480	15.439 ± 5.990	0.209 ± 0.042
	3 T	91.250 ± 1.461	**99.995 ± 0.023**	97.851 ± 0.786	15.736 ± 5.910	0.218 ± 0.051
MBWSS	1.5 T	97.583 ± 2.214	95.765 ± 3.198	**99.940 ± 0.167**	22.866 ± 7.930	0.069 ± 0.273
	3 T	96.233 ± 5.421	93.506 ± 8.408	**99.959 ± 0.050**	27.791 ± 9.984	0.248 ± 0.830
OPTIBET	1.5 T	96.534 ± 0.575	97.239 ± 0.650	99.471 ± 0.165	12.542 ± 10.222	0.053 ± 0.012
	3 T	96.595 ± 0.813	97.578 ± 0.747	99.484 ± 0.343	12.139 ± 7.854	0.055 ± 0.019
ROBEX	1.5 T	95.714 ± 0.953	98.798 ± 0.562	99.033 ± 0.273	10.727 ± 3.495	0.077 ± 0.045
	3 T	95.782 ± 1.122	98.583 ± 0.792	99.207 ± 0.352	10.391 ± 3.291	0.078 ± 0.052
STAPLE	1.5 T	97.990 ± 0.852	99.801 ± 0.165	99.522 ± 0.209	**6.770 ± 2.099**	**0.014 ± 0.004**
	3 T	97.325 ± 1.146	99.919 ± 0.107	99.428 ± 0.267	**6.883 ± 2.040**	**0.014 ± 0.013**

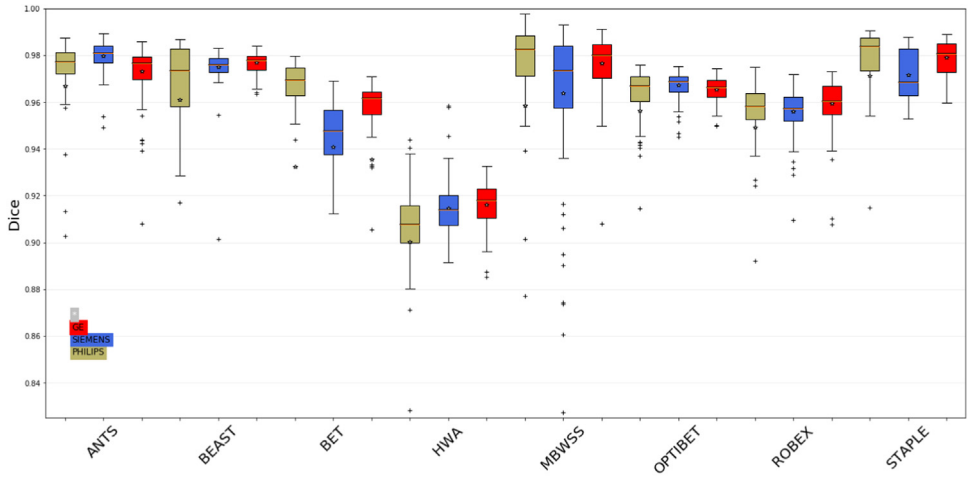
distributions were different ($p - value < 0.001$). No significant gender distribution differences were found between the six scanner/field strength groups ($p = 0.290$, Chi-squared test).

3.2. Skull stripping and comparison against "silver standard" consensus mask

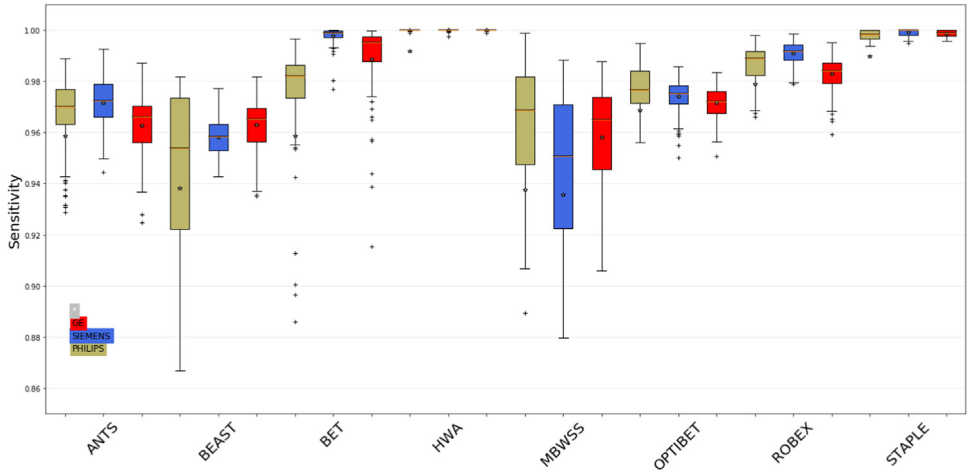
The nine different skull stripping results were evaluated against the consensus "silver-standard" to assess their agreement. The overall metrics are summarized in Table 2. Statistical significance of the agreement between methods was computed pairwise using the Wilcoxon signed-rank test with Bonferroni correction. The results versus the classifier ("silver standard") are summarized by the elements above the main diagonal in Table 3. The analyses by scanner vendor and magnetic field strength are summarized in Tables 4 and 5, respectively. Dice, sensitivity and specificity box-plots grouped by vendor and magnetic field strength are depicted in Figs. 2 and 3, respectively. Table 6 summarizes the vendor-magnetic field strength interaction.

3.3. Influence of vendor and magnetic field strength

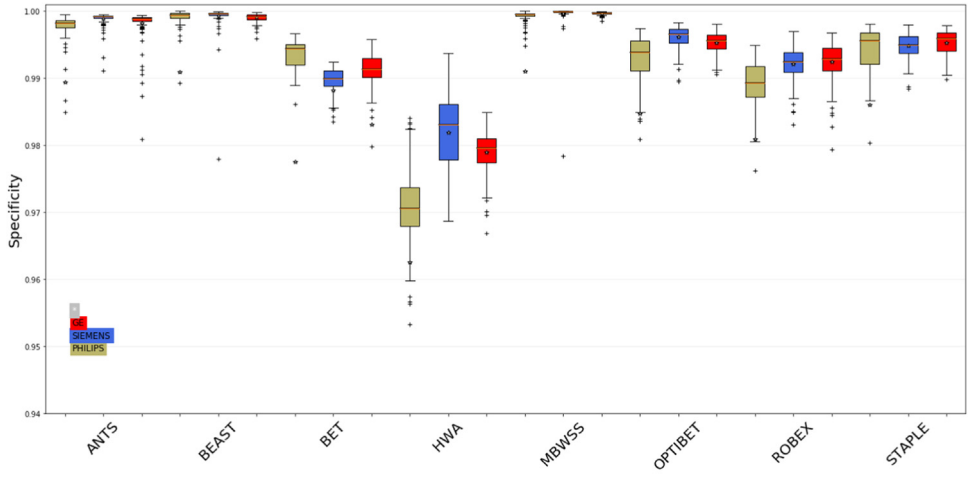
A LME model of the Dice coefficient results was constructed to assess gender, scanner vendor and magnetic field strength influences on skull stripping. We removed BET and BSE from this model, because they had high variances, which would violate the



(a) Dice



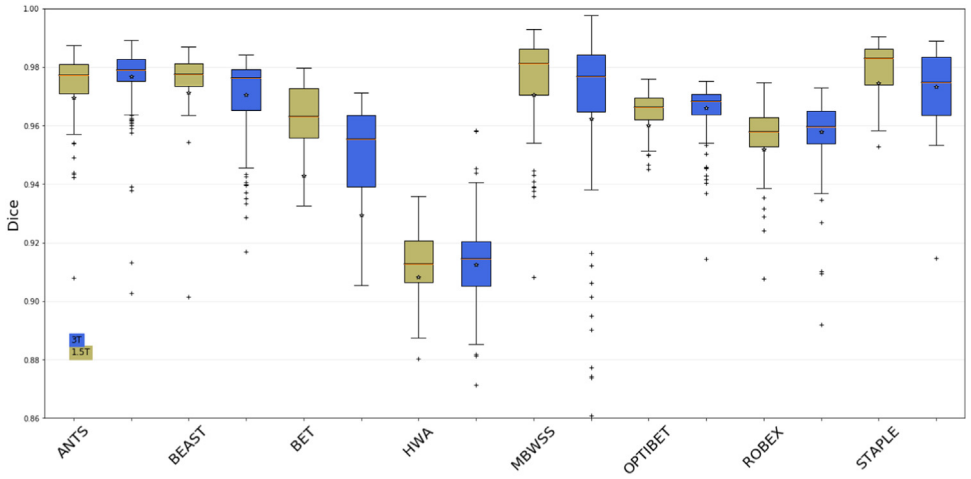
(b) Sensitivity



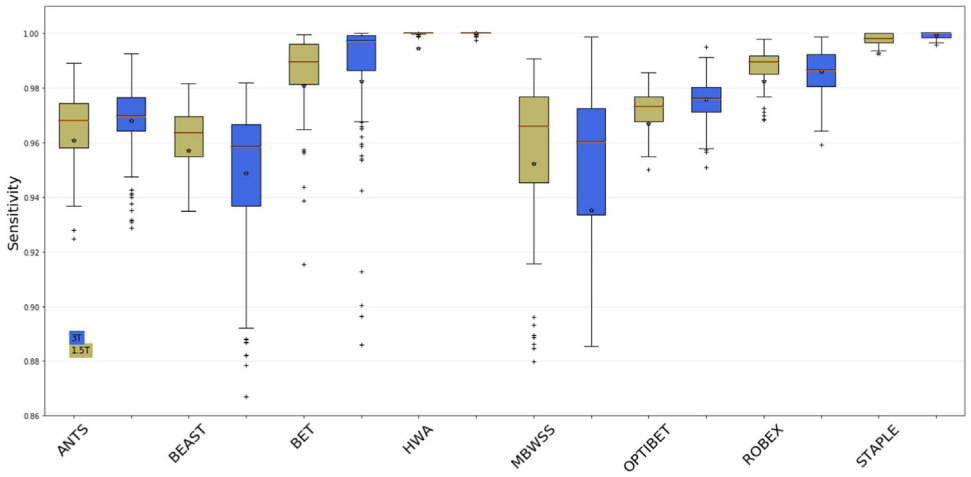
(c) Specificity

Fig. 2. Average Dice coefficient, sensitivity and specificity metric by vendor. BSE results were excluded for better scaling of the data.

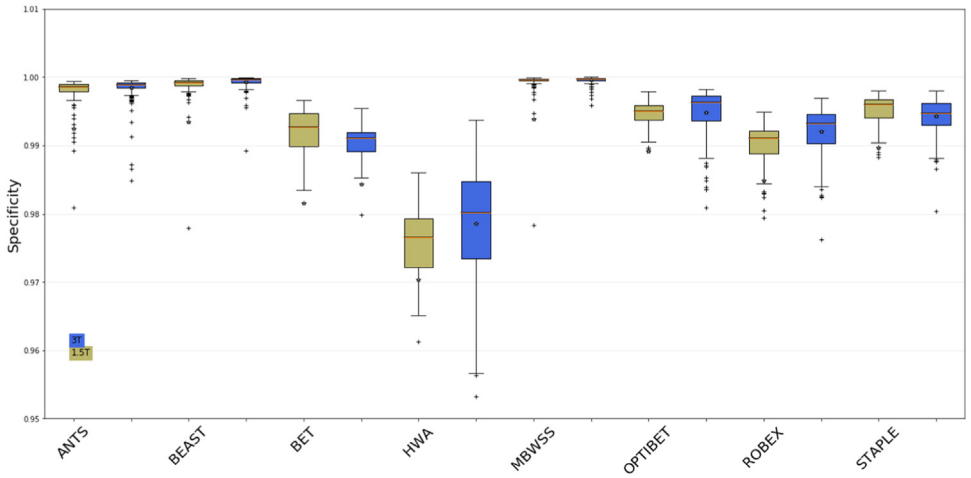
model assumptions. The residual plots, the residual histogram and the residual-symmetry plot (or Q-Q plot; Fig. 4) confirmed that the linearity, homoscedasticity and the residual normality assumptions of the LME model after removal of BET and BSE from the model were met. The coefficients, standard errors, p -values and confidence intervals of the model are summarized in Table 8.



(a) Dice



(b) Sensitivity



(c) Specificity

Fig. 3. Average Dice coefficient, sensitivity and specificity metric by magnetic field strength. BSE results were excluded for better scaling of the data.

3.4. *Comparison against manual segmentation results*

We compared the eight skull stripping techniques and both consensus methods against the twelve manual segmentation results. The STAPLE mask was thresholded at probability 0.5. The "silver standard" consensus results are the average of a 2-fold

Table 6
Summary of vendor-magnetic field strength interaction (Dice, sensitivity, specificity, Hausdorff distance and symmetric distance). The best score for each scanner vendor-field intensity combination for each metric are emboldened.

Method	Vendor	Field	Dice	Sensitivity	Specificity	Hausdorff	Mean distance
ANTS	Philips	1.5	98.084 ± 0.381	97.589 ± 0.571	99.791 ± 0.092	8.573 ± 1.623	0.027 ± 0.009
		3.0	96.903 ± 1.356	95.753 ± 2.140	99.743 ± 0.251	7.829 ± 2.751	0.051 ± 0.059
	Siemens	1.5	97.588 ± 0.656	96.540 ± 0.760	99.847 ± 0.136	9.376 ± 8.581	0.037 ± 0.025
		3.0	**98.338 ± 0.290**	97.768 ± 0.541	99.919 ± 0.024	8.766 ± 1.794	0.024 ± 0.006
	GE	1.5	96.843 ± 1.181	95.702 ± 1.197	99.772 ± 0.304	8.495 ± 2.358	0.055 ± 0.053
		3.0	97.775 ± 0.630	96.851 ± 0.623	99.854 ± 0.165	9.592 ± 1.979	0.036 ± 0.027
BEaST	Philips	1.5	98.248 ± 0.239	97.271 ± 0.515	99.890 ± 0.062	8.435 ± 1.634	0.024 ± 0.006
		3.0	95.564 ± 1.199	91.963 ± 2.131	99.938 ± 0.147	8.908 ± 1.784	0.070 ± 0.028
	Siemens	1.5	97.305 ± 1.015	95.659 ± 0.653	99.883 ± 0.285	10.600 ± 10.487	0.046 ± 0.071
		3.0	97.698 ± 0.325	95.937 ± 0.764	99.965 ± 0.021	8.974 ± 1.874	0.034 ± 0.007
	GE	1.5	97.453 ± 0.417	95.809 ± 0.996	99.906 ± 0.053	8.295 ± 2.249	0.036 ± 0.008
		3.0	97.890 ± 0.343	96.758 ± 0.757	99.894 ± 0.066	12.457 ± 16.945	0.034 ± 0.009
BET	Philips	1.5	97.457 ± 0.303	98.238 ± 0.640	99.503 ± 0.067	9.060 ± 1.804	0.034 ± 0.009
		3.0	90.625 ± 15.491	95.128 ± 12.605	97.657 ± 3.809	24.346 ± 29.924	1.930 ± 6.419
	Siemens	1.5	94.996 ± 4.160	99.634 ± 0.286	98.781 ± 1.249	13.783 ± 11.269	0.232 ± 1.120
		3.0	93.132 ± 4.606	99.908 ± 0.294	98.861 ± 1.026	20.451 ± 16.097	0.404 ± 1.919
	GE	1.5	92.031 ± 11.461	98.010 ± 2.512	97.839 ± 4.068	21.248 ± 29.100	1.207 ± 3.257
		3.0	95.081 ± 5.459	99.694 ± 0.164	98.775 ± 2.032	14.790 ± 12.991	0.233 ± 1.164
BSE	Philips	1.5	95.132 ± 1.968	92.066 ± 3.513	99.789 ± 0.148	20.504 ± 12.589	0.148 ± 0.142
		3.0	89.703 ± 12.668	85.929 ± 11.553	99.011 ± 3.197	58.321 ± 20.595	1.239 ± 5.490
	Siemens	1.5	92.971 ± 17.300	91.764 ± 17.175	99.404 ± 1.659	56.237 ± 19.122	1.201 ± 5.848
		3.0	95.070 ± 1.715	91.892 ± 3.271	99.897 ± 0.031	51.932 ± 19.336	0.188 ± 0.090
	GE	1.5	92.484 ± 3.760	87.981 ± 7.102	99.772 ± 0.109	41.093 ± 24.945	0.302 ± 0.237
		3.0	75.116 ± 26.082	91.590 ± 24.488	92.579 ± 7.228	100.561 ± 22.513	8.255 ± 10.020
HWA	Philips	1.5	91.069 ± 0.954	**99.999 ± 0.002**	97.102 ± 0.340	14.986 ± 1.659	0.217 ± 0.042
		3.0	90.481 ± 1.851	**99.995 ± 0.017**	97.024 ± 0.670	15.751 ± 1.758	0.237 ± 0.063
	Siemens	1.5	91.322 ± 0.975	**99.997 ± 0.004**	97.759 ± 0.332	16.232 ± 10.063	0.209 ± 0.045
		3.0	91.599 ± 1.173	**99.991 ± 0.034**	98.612 ± 0.242	15.882 ± 1.836	0.219 ± 0.043
	GE	1.5	91.555 ± 0.933	**99.989 ± 0.026**	97.864 ± 0.356	15.092 ± 1.455	0.200 ± 0.035
		3.0	91.671 ± 0.844	**100.000 ± 0.000**	97.918 ± 0.281	15.576 ± 9.913	0.198 ± 0.035
MBWSS	Philips	1.5	98.077 ± 1.016	96.773 ± 1.899	**99.919 ± 0.077**	26.055 ± 6.575	0.043 ± 0.038
		3.0	95.255 ± 8.447	92.380 ± 12.693	**99.937 ± 0.068**	32.045 ± 12.405	0.469 ± 1.355
	Siemens	1.5	97.168 ± 3.384	95.074 ± 4.376	**99.936 ± 0.275**	19.285 ± 8.945	0.106 ± 0.451
		3.0	95.616 ± 3.478	92.013 ± 6.135	**99.983 ± 0.033**	26.875 ± 8.675	0.234 ± 0.371
	GE	1.5	97.513 ± 1.321	95.465 ± 2.498	**99.965 ± 0.018**	23.309 ± 6.487	0.059 ± 0.127
		3.0	97.828 ± 0.899	96.125 ± 1.720	**99.956 ± 0.025**	24.454 ± 6.312	0.041 ± 0.016
OPTIBET	Philips	1.5	96.893 ± 0.422	97.190 ± 0.613	99.495 ± 0.156	10.968 ± 1.713	0.045 ± 0.007
		3.0	95.970 ± 1.078	98.153 ± 0.622	99.106 ± 0.349	11.929 ± 6.429	0.068 ± 0.026
	Siemens	1.5	96.497 ± 0.598	97.216 ± 0.739	99.495 ± 0.164	13.595 ± 12.071	0.055 ± 0.014
		3.0	96.959 ± 0.327	97.582 ± 0.483	99.720 ± 0.064	13.210 ± 9.948	0.050 ± 0.007
	GE	1.5	96.218 ± 0.476	97.311 ± 0.579	99.424 ± 0.164	13.037 ± 12.622	0.059 ± 0.010
		3.0	96.856 ± 0.350	96.999 ± 0.622	99.625 ± 0.098	11.276 ± 6.543	0.048 ± 0.009
ROBEX	Philips	1.5	96.214 ± 0.900	99.062 ± 0.497	98.976 ± 0.313	9.660 ± 3.074	0.063 ± 0.038
		3.0	95.227 ± 1.123	98.343 ± 0.646	98.849 ± 0.327	10.526 ± 1.979	0.087 ± 0.031
	Siemens	1.5	95.655 ± 0.860	98.846 ± 0.401	99.078 ± 0.233	11.329 ± 4.063	0.080 ± 0.036
		3.0	95.547 ± 0.924	99.350 ± 0.276	99.346 ± 0.173	10.729 ± 2.790	0.086 ± 0.044
	GE	1.5	95.282 ± 0.858	98.490 ± 0.611	99.044 ± 0.257	11.173 ± 2.994	0.090 ± 0.054
		3.0	96.572 ± 0.821	98.056 ± 0.683	99.425 ± 0.196	9.919 ± 4.522	0.061 ± 0.070
STAPLE	Philips	1.5	**98.696 ± 0.372**	99.648 ± 0.121	99.662 ± 0.123	**7.929 ± 1.747**	**0.014 ± 0.003**
		3.0	**97.172 ± 1.173**	99.961 ± 0.072	99.200 ± 0.296	**5.720 ± 1.902**	**0.017 ± 0.022**
	Siemens	1.5	**97.826 ± 0.855**	99.837 ± 0.119	99.497 ± 0.217	**5.984 ± 1.924**	**0.012 ± 0.003**
		3.0	96.485 ± 0.800	99.976 ± 0.071	99.449 ± 0.147	**6.865 ± 1.495**	**0.011 ± 0.002**
	GE	1.5	97.460 ± 0.707	99.914 ± 0.126	99.409 ± 0.189	**6.417 ± 2.094**	**0.014 ± 0.004**
		3.0	**98.318 ± 0.458**	99.820 ± 0.095	99.634 ± 0.098	**8.064 ± 1.971**	**0.013 ± 0.003**

Table 7
Overall outlier analysis. Number (percentage) of outlying segmentations for each evaluated method versus the "silver standard" consensus mask. Evaluated at Dice coefficients thresholds of 0.8, 0.85 and 0.9.

Method	Dice <80%	Dice <85%	Dice <90%
ANTs	0 (0.0%)	0 (0.0%)	0 (0.0%)
BEaST	0 (0.0%)	0 (0.0%)	0 (0.0%)
BET	19 (5.3%)	19 (5.3%)	19 (5.3%)
BSE	35 (9.8%)	39 (10.9%)	60 (16.7%)
HWA	0 (0.0%)	1 (0.3%)	39 (10.9%)
MBWSS	6 (1.7%)	7 (1.9%)	13 (3.6%)
OPTIBET	0 (0.0%)	0 (0.0%)	0 (0.0%)
ROBEX	0 (0.0%)	0 (0.0%)	1 (0.3%)
STAPLE	0 (0.0%)	0 (0.0%)	0 (0.0%)

Table 8
LME model output for Dice coefficient as a function of scanner vendor, magnetic field strength and gender as fixed effects. Significant p -values are highlighted emboldened.

Coefficient	Estimate	Standard error	p-value	95% CI
(Intercept)	9.590e − 01	9.773e − 03	**6.950e-10**	(0.938, 0.980)
Field	−2.841e − 03	8.495e − 04	**0.001**	(−0.005, − 0.001)
Vendor	2.416e − 03	5.320e − 04	**6.460e-06**	(0.001, 0.003)
Gender	−1.100e − 03	8.692e − 04	0.206	(−0.003, 0.001)

cross-validation. Each fold was trained with six subjects, one for each scanner vendor and magnetic field intensity combination. Each fold had approximately six million training samples (voxels) and achieved training accuracy of 94.0% (fold 1) and 93.9% (fold 2). *p* -Values were computed (Wilcoxon signed-rank test with Bonferroni correction) using the "silver standard" consensus as the reference method. The global metrics are summarized in Table 9 and the statistical significance results are summarized by the values below the main diagonal in Table 3.

Receiver-operating characteristic (ROC) curve was generated for the STAPLE method at different probability thresholds. The area under the ROC curve was 0.99. The Dice coefficient was also plotted against the STAPLE probability threshold (Fig. 5). A maximum Dice coefficient was found at a threshold of probability <0.99.

4. Discussion

The overall analysis of the skull stripping techniques against the "silver-standard" consensus showed that STAPLE, ANTS and BEaST achieved the highest Dice coefficient metrics and had smaller variance (standard deviation), therefore they have high agreement with the consensus mask and their performance was consistent. STAPLE's Dice coefficient was significantly different compared to all the methods, except for ANTs and MBWSS.

The Dice coefficient metric represents a compromise between sensitivity (including brain tissue) and specificity (not including non-brain tissue). STAPLE, OPTIBET, BEaST and ANTS were found to be robust techniques; their Dice coefficients were higher than 0.9 for all 359 subjects assessed (Table 7). BSE, BET are the less consistent (robust) with higher standard deviations. MBWSS has the fourth highest Dice coefficient, it suffers from failing in some cases (13 subjects; Dice <0.9) by leaving a big portion of the brain out of the segmentation mask. If we excluded these failures, MBWSS would have an average Dice coefficient of 97.58 ± 1.49, which is close to the results obtained by ANTs and STAPLE.

STAPLE and HWA were very sensitive methods and included most brain tissue, a feature that may be very important in some applications. Nevertheless, HWA achieved a high sensitivity at the cost of reduced specificity; it has the lowest specificity among the

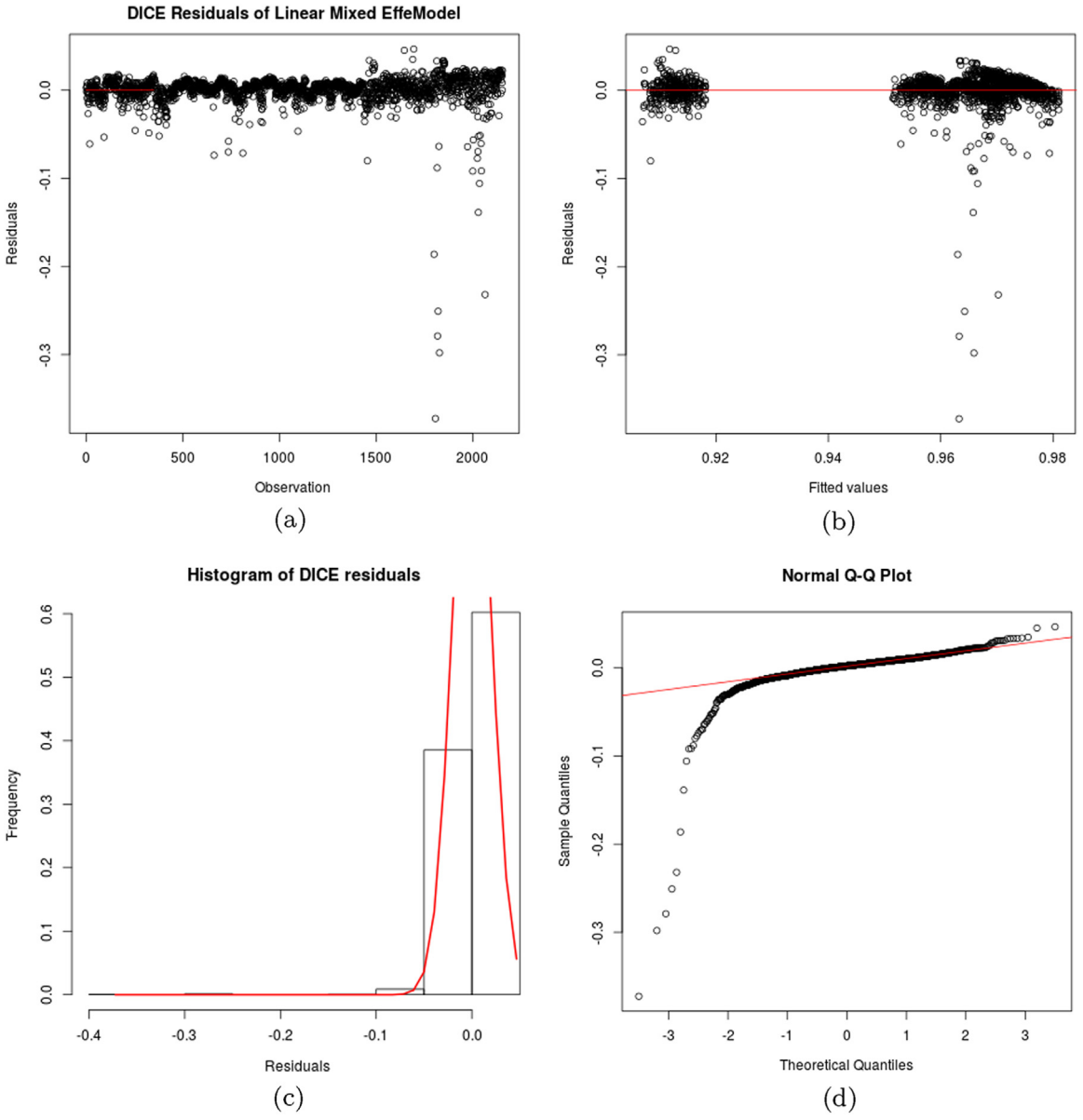


Fig. 4. LME model assumptions verification. (a) Residuals plot. (b) Residuals versus fitted values plot. (c) Histogram of residuals. (d) Residuals symmetry plot.

Table 9
Overall analysis against manual segmentation results (Dice, sensitivity, specificity, Hausdorff distance and mean symmetric distance). The two best scores for each metric are emboldened.

Method	Dice	Sensitivity	Specificity	Hausdorff	Mean distance
ANTS	95.927 ± 0.872	94.510 ± 1.583	99.705 ± 0.114	8.905 ± 1.393	0.057 ± 0.015
BEaST	95.766 ± 1.225	93.838 ± 2.568	**99.757 ± 0.133**	9.907 ± 1.410	0.067 ± 0.029
BET	95.220 ± 0.937	98.261 ± 1.610	99.131 ± 0.232	12.169 ± 2.766	0.080 ± 0.024
BSE	90.488 ± 7.028	91.441 ± 5.319	98.648 ± 2.267	61.416 ± 29.040	1.562 ± 3.179
HWA	91.657 ± 1.110	**99.930 ± 0.122**	97.830 ± 0.824	15.399 ± 1.799	0.179 ± 0.038
MBWSS	95.568 ± 1.455	92.784 ± 2.668	**99.848 ± 0.039**	28.228 ± 5.446	0.080 ± 0.031
OPTIBET	95.433 ± 0.705	96.133 ± 0.952	99.357 ± 0.305	10.304 ± 1.998	0.066 ± 0.013
ROBEX	95.611 ± 0.724	98.421 ± 0.703	99.130 ± 0.281	9.410 ± 1.610	0.063 ± 0.015
STAPLE	**96.797 ± 0.744**	**98.976 ± 0.596**	99.382 ± 0.220	**8.327 ± 1.665**	**0.038 ± 0.007**
"Silver Standard"	**97.135 ± 0.511**	96.825 ± 0.677	99.709 ± 0.108	**7.952 ± 0.888**	**0.036 ± 0.007**

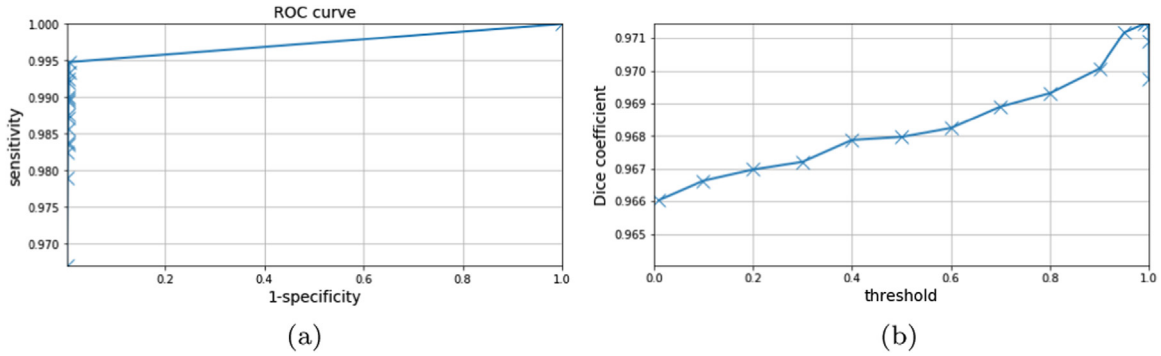


Fig. 5. (a) STAPLE receiver-operator characteristic (ROC) curve for varying STAPLE probability thresholds, and (b) Dice coefficient plotted against STAPLE probability threshold. Area under the ROC curve was 0.99.

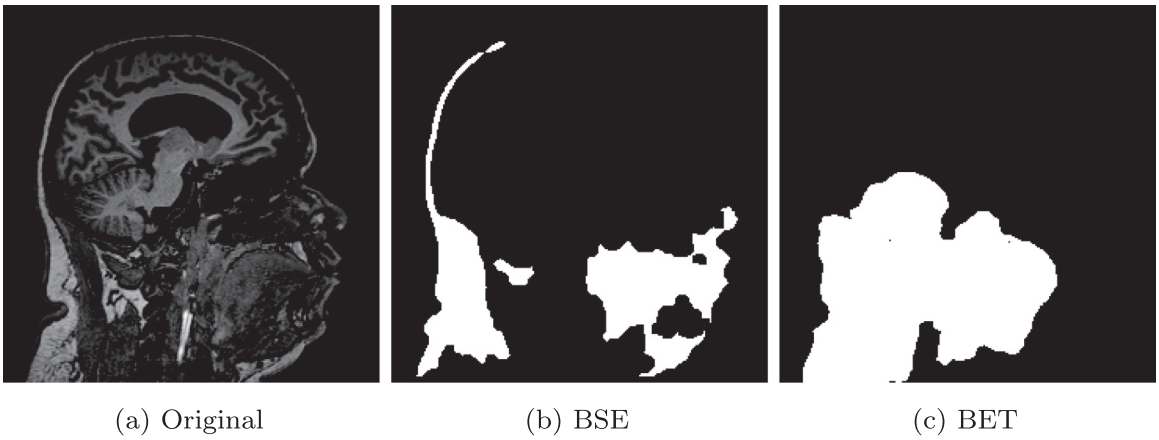


Fig. 6. (a) Central sagittal slice of a Philips 3 T T1 volume. Two examples of segmentation failures: (b) BSE method segments the skull instead of the brain. (c) BET method segments the neck region instead of the brain.

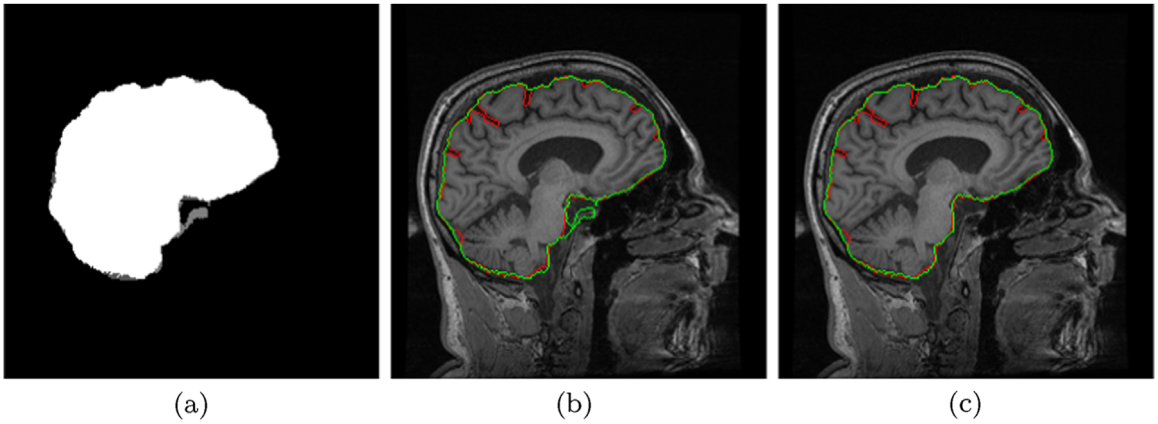


Fig. 7. (a) STAPLE brain mask probability map. Manual segmentation (red line) and STAPLE brain mask thresholded at probability: (b) >0.5 (green line) and (c) probability >0.95 (green line).

assessed techniques. ANTs, BEaST and MBWSS were the most specific methods and nearly always failed to include non-brain tissue in the final brain segmentation mask.

BSE achieved the poorest results, especially on Philips 3 T and GE 3 T data. The BSE method completely failed to segment the brain in seven subjects. In all of these cases, it segmented the skull instead of the brain. This same skull stripping error also occurred in one subject with BET (see Fig. 6).

These results showed that some techniques, such as MBWSS, BSE, and BET, performed better on 1.5 T rather than 3 T data (Tables 5 and 6; Fig. 3). Although 3 T scanners have increased signal-to-noise ratio (SNR), most skull stripping techniques were initially developed when 1.5 T was the predominant field intensity for brain imaging. This fact likely provides an explanation for why most techniques performed better at 1.5 T. The increased susceptibility effects observed in more inferior slices at 3 T may also contribute to this finding. Scanner vendor also was found to influence skull stripping performance (Table 8). This effect can potentially be explained by the use of different, vendor-specific, reconstruction and image filtering algorithms.

The ROC curve for the STAPLE results (Fig. 5) indicates a behavior close to ideal (area under the ROC curve of 0.99) and that the STAPLE probability mask is close to binary with a few exceptions as illustrated in Fig. 7. We report only STAPLE's ROC, because in order to build the curve it is necessary for a parameter to be thresholded. Some methods have more than one parameter and

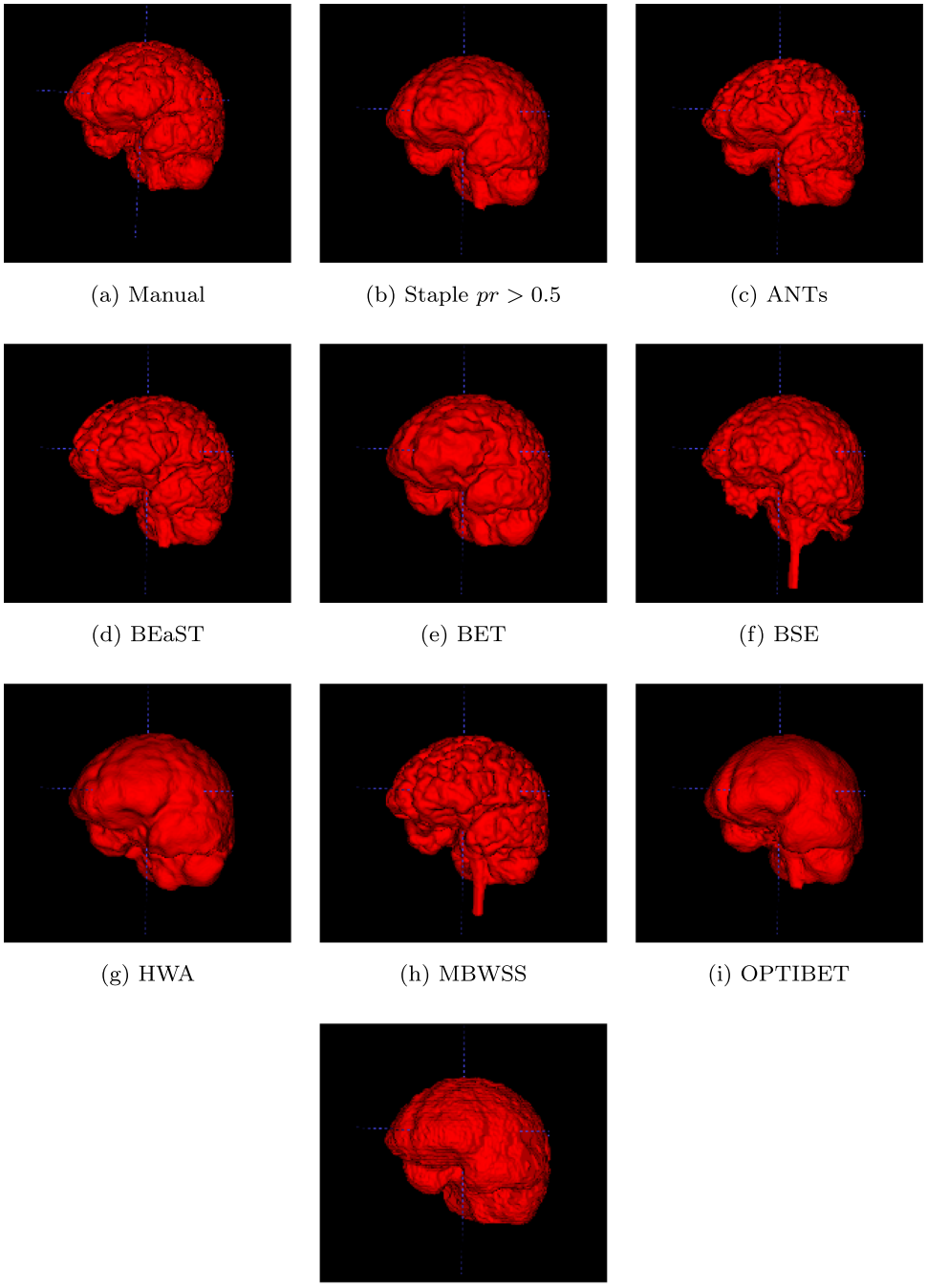


Fig. 8. Representative 3D reconstruction of the different segmentation methods for one subject on a GE scanner at 1.5 T.

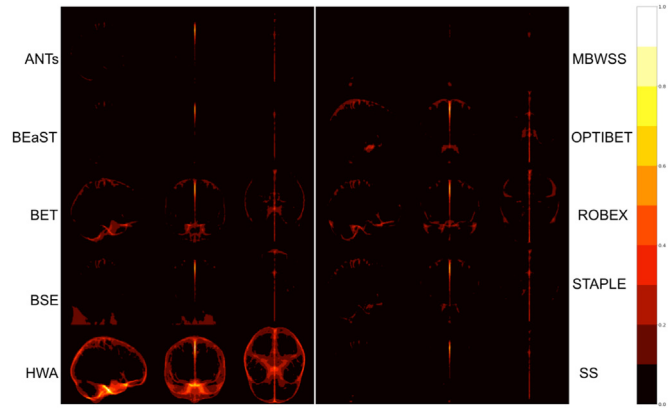


Fig. 9. Sagittal, coronal and axial heat map projections of FP using the manual segmentations as reference.

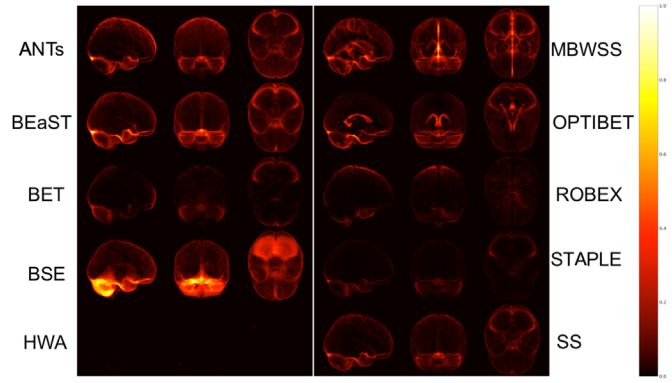


Fig. 10. Sagittal, coronal and axial heat map projections of FN using the manual segmentations as reference.

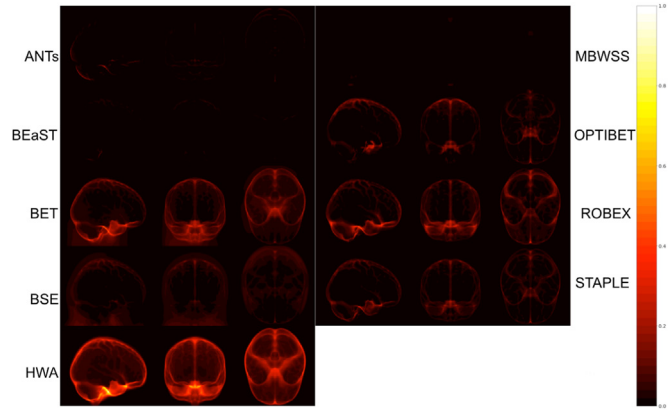


Fig. 11. Sagittal, coronal and axial heat map projections of FP using the "silver-standards" as reference.

others have none, making it harder to perform ROC analysis for the other methods.

BSE and BET were not considered in the LME model, because they exhibited much larger variability compared to the other skull stripping methods (Table 2), which would have violated LME model assumptions. The results of the LME model confirmed that magnetic field strength and vendor influence brain extraction. Subject gender did not influence skull stripping ($p = 0.206$). Age was modeled as a random effect, because we found that there was a significant statistical difference in the age distribution by scanner vendor/field strength.

In the comparison against the twelve manual segmentations, the consensus obtained using the logistic regression classifier achieved the highest Dice coefficient. The difference was statistically significant

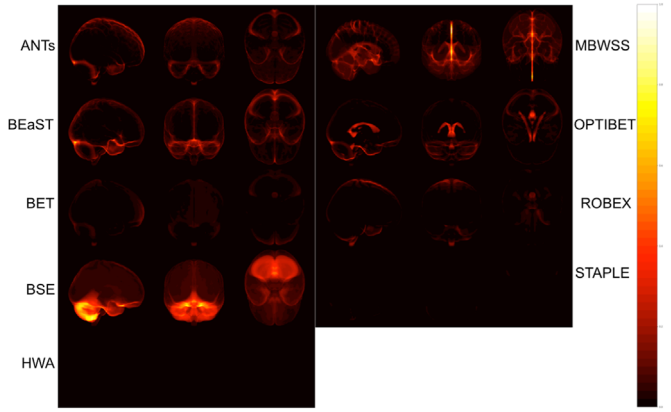


Fig. 12. Sagittal, coronal and axial heat map projections of FN using the "silver-standards" as reference.

compared to all methods, except STAPLE (Table 3). The logistic regression classifier also achieved the smallest average Haussdorff distance (statistical significant except versus STAPLE) and the smallest symmetric surface-to-surface mean distance (statistical significant except versus ANTs and STAPLE). STAPLE achieved the second highest Dice coefficient and the second highest sensitivity. It also achieved the second smallest Haussdorff distance and symmetric surface-to-surface mean distance. These results are an indicative that consensus approaches improve the performance of the individual methods used to generate the consensus.

It is interesting to note that most evaluated segmentation techniques were not able to properly follow the complexities of the brain cortical surface (Fig. 8). OPTIBET produces a smooth surface and ROBEX has a mosaicking aspect. MBWSS and BEaST are the ones that seem to follow better he complexities of the brain cortical surface. Also, MBWSS and BSE tended to preserve the spinal cord in their segmentation mask, which is not incorrect. However this inclusion degraded their performance metrics, specially the Haussdorff distance, due the fact that the comparative manual segmentation masks do not include these structures.

We used the non-linear registration implemented in (Avants et al., 2011) to take all subjects to the same space. The subjects were registered using a symmetric atlas. Then, we computed the false positive (FP) and false negative (FN) average error projection for all the skull stripping methods using the manually segmented subjects as reference. Sagittal, coronal and axial false positive and false negative error projections are shown as heat maps in Figs. 9 and 10, respectively. The heat maps were normalized between 0 and 1. The upper extreme represents a high systematic number of FPs (Fig. 9) and FNs (Fig. 10). Fig. 9 shows that even methods with high specificity, such as ANTs and BEaST, were not able to properly segment the brain fissure between the left and right brain hemispheres. Among the methods evaluated, only MBWSS was capable of correctly segmenting the brain fissure.

Figs. 11 and 12 also depict heat maps, but using the "silver standards" as reference. We can see BET and BSE include the neck in some of their results. Also, since seven out of the eight methods used to generate the "silver standard" incorrectly include the brain fissure in their segmentation mask, the "silver-standard" masks also include it in their segmentation as being brain, which is an additional reason for calling the metrics computed against the "silver-standard" agreement and not accuracy.

The heat maps depicted are fairly symmetric mainly because the skull stripping methods being assessed do not distinguish between left and right brain hemispheres. Therefore, they make errors on both hemispheres. Since we are using a symmetric reference atlas and the projections represent averages, the error distributions look consistently symmetric across hemispheres for

all algorithms, but closer examination shows that there are some small asymmetries.

In summary, ANTs and BEaST were the best performing techniques in terms of Dice coefficient and also robustness (small standard deviations; Dice >90% for all subjects). Nevertheless, MBWSS achieves close to average results, despite being less robust. Also, MBWSS was the only method capable of segmenting properly the brain fissure and its comparison had a small disadvantage compared to the others, since it preserves the spinal cord while the other methods (except BSE) do not preserve the spinal cord. By manually removing the spinal cord from the MBWSS segmentation masks, the Dice coefficient could be increased on average by 0.5%. MBWSS processing time is in the order of seconds, while ANTs and BEaST take a few minutes to process.

5. Conclusions

We have proposed and developed a public, multi-centre, multi-field strength T1 3D brain MR dataset and used it to evaluate agreement between eight publicly available skull stripping techniques plus the STAPLE algorithm. The overall analysis indicated that STAPLE, ANTS and BEaST achieve the best Dice coefficients, which reflects a compromise between sensitivity and specificity. Also, although not as robust as ANTs and BEaST, MBWSS obtains comparable results and is capable of correctly segmenting the brain fissure. Methods like HWA are extremely sensitive and do not exclude brain tissue from the segmentation mask; while methods like ANTs, MBWSS and BEaST are more specific. BET and BSE were two methods with high variance, therefore judged to be less consistent in their segmentations. The LME analysis indicated that the scanner vendor and the magnetic field strength have significant influence on the skull stripping results.

The choice of brain extraction method is problem dependent and is influenced by characteristics of the MR images. Factors, such as scanner vendor and magnetic field intensity, should be considered when selecting the most appropriate skull stripping method. The CC-359 dataset can be used to evaluate and/or optimize skull stripping parameters by vendor and magnetic field strength. The consensus masks can be used as labeled data for a number of tasks including training deep neural networks (Kleesiek et al., 2016).

Acknowledgments

This project was supported by FAPESP CEPID-BRAINN (2013/07559-3) and CAPES PVE (88881.062158/2014-01). Roberto A. Lotufo thanks CNPq (311228/2014-3), Simone Appenzeller thanks CNPq (157534/2015-4), Roberto Souza thanks FAPESP (2013/23514-0) and the NSERC CREATE I3T foundation, Oeslle Lucena thanks FAPESP (2016/18332-8). Richard Frayne is supported by the Canadian Institutes of Health Research (CIHR, MOP-333931) and the Hopewell Professorship in Brain Imaging. Infrastructure in the Calgary Image Processing and Analysis Centre (CIPAC) was partially developed with funding provided by the Canada Foundation for Innovation and the Government of Alberta.

References

Asman, A.J., Landman, B.A., 2011. Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE). IEEE Trans. Med. Imaging 30 (10), 1779–1794.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage 54 (3), 2033–2044.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48.

R. Beare, J. Chen, C. Adamson, T. Silk, D. Thompson, J. Yang, V. Anderson, M. Seal, A. Wood, Brain extraction using the watershed transform from markers, Front. Neuroinformatics 7(32).

Boesen, K., Rehm, K., Schaper, K., Stoltzner, S., Woods, R., Lüders, E., Rottenberg, D., 2004. Quantitative comparison of four brain extraction algorithms. NeuroImage 22 (3), 1255–1261.

Bouix, S., Martin-Fernandez, M., Ungar, L., Nakamura, M., Koo, M.-S., McCarley, R.W., Shenton, M.E., 2007. On evaluating brain tissue classifiers without a ground truth. NeuroImage 36 (4), 1207–1224.

Breiman, L., 2001. Random forests. Mach. Learn 45 (1), 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees, first ed. Wadsworth International Group.

Collins, M., Schapire, R.E., Singer, Y., 2002. Logistic regression, AdaBoost and Bregman distances. Mach. Learn. 48 (1), 253–285.

Cox, R.W., Ashburner, J., Breman, H., Fissell, K., Haselgrove, C., Holmes, C.J., Lancaster, J.L., Rex, D.E., Smith, S.M., Woodward, J.B., Strother, S.C., 2004. A (sort of) new image data format standard: NIfTI-1, Tenth Annual Meeting of the Organization for Human Brain Mapping.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. NeuroImage, 179–194.

Eskildsen, S.F., Coupé, P., Fonov, V., Manjòn, J.V., Leung, K.K., Guizard, N., Wassef, S. N., Østergaard, L.R., Collins, D.L., 2012. BEaST: brain extraction based on nonlocal segmentation technique. NeuroImage 59 (3), 2362–2373.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Statistics, 1189–1232.

Hahn, H.K., Peitgen, H., 2000. The skull stripping problem in MRI solved by a single 3D watershed transform, Proceedings of the Third International Conference on Medical Image Computing and Computer-assisted Intervention, MICCAI '00. Springer-Verlag, London, UK, UK, pp. 134–143.

Haynes, W., 2013. Wilcoxon Rank Sum Test. Springer New York, New York, NY, pp. 2354–2355.

Ibáñez, L., Schroeder, W., Ng, L., Cates, J., 2003. The ITK Software Guide. Kitware.

Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging 30 (9), 1617–1634.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. FSL. NeuroImage 62 (2), 782–790.

Khastavaneh, H., Ebrahimpour-Komleh, H., 2015. Brain extraction: a region based histogram analysis strategy, 2015 Signal Processing and Intelligent Systems Conference (SPIS), pp. 20–24.

Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., Biller, A., 2016. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. NeuroImage 129, 460–469.

Lee, J., Yoon, U., Nam, S., Kim, J., Kim, I., Kim, S., 2003. Evaluation of automated and semi-automated skull-stripping algorithms using similarity index and segmentation error. Comput. Biol. Med. 33 (6), 495–507.

Lutkenhoff, E.S., Rosenberg, M., Chiang, J., Zhang, K., Pickard, J.D., Owen, A.M., Monti, M.M., 2014. Optimized brain extraction for pathological brains (OPTI-BET). PLoS ONE 9 (12), 1–13.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19 (9), 1498–1507.

Rehm, K., Schaper, K., Anderson, J., Woods, R., Stoltzner, S., Rottenberg, D., 2004. Putting our heads together: a consensus approach to brain/non-brain segmentation in T1-weighted MR volumes. NeuroImage 22 (3), 1262–1270.

Rex, D.E., Shattuck, D.W., Woods, R.P., Narr, K.L., Luders, E., Rehm, K., Stolzner, S.E., Rottenberg, D.A., Toga, A.W., 2004. A meta-algorithm for brain extraction in MRI. NeuroImage 23 (2), 625–637.

Roy, S., Maji, P., 2015. A simple skull stripping algorithm for brain MRI, 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR), pp. 1–6.

F. Ségonne, A. M. Dale, B. E. Busa, B. M. Glessner, B. D. Salat, B. H. K. Hahn, B. F. A, A hybrid approach to the skull stripping problem in MRI, NeuroImage 22.

Shattuck, D.W., Leahy, R.M., 2000. BrainSuite: an Automated Cortical Surface Identification Tool. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 50–61.

Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., 2001. Magnetic resonance image tissue classification using a partial volume model. NeuroImage 13 (5), 856–876.

Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W., 2008. Construction of a 3D probabilistic atlas of human cortical structures. NeuroImage 39 (3), 1064–1080.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155.

Soille, P., 2004. Erosion and Dilation. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 63–103.

Steinwart, I., Christmann, A., 2008. Support Vector Machines, first ed. Springer Publishing Company, Incorporated.

Wang, H., Yushkevich, P., 2013. Multi-atlas segmentation with joint label fusion and corrective learning — an open source implementation. Front. Neuroinformatics 7, 27.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921.

Yushkevich, P.A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage 31 (3), 1116–1128.