Breast MRI segmentation for density estimation: Do different methods give the same results and how much do differences matter?

Simon J. Doran,[1, a)] John H. Hipwell,[2, b)] Rachel Denholm,[3] Björn Eiben,[2] Marta Busana,[3] David J. Hawkes,[2] Martin O. Leach,[1] and Isabel dos Santos Silva[3]

[1)] *Cancer Research UK Cancer Imaging Centre, Division of Radiotherapy and Imaging, The Institute of Cancer Research, London, SM2 5NG, UK.*

[2)] *Centre for Medical Image Computing (CMIC), Department of Medical Physics and Bioengineering, UCL, London, WC1E 7JE, UK.*

[3)] *Department of Non-Communicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK.*

(Dated: 1 March 2017)

ABSTRACT

Purpose To compare two methods of automatic breast segmentation with each other and with manual segmentation in a large subject cohort. To discuss the factors involved in selecting the most appropriate algorithm for automatic segmentation and, in particular, to investigate the appropriateness of overlap measures (e.g., Dice and Jaccard coefficients) as the primary determinant in algorithm selection.

Methods Two methods of breast segmentation were applied to the task of calculating MRI breast density in 200 subjects drawn from the Avon Longitudinal Study of Parents and Children, a large cohort study with an MRI component.

A semi-automated, bias-corrected, fuzzy C-means (BC-FCM) method was combined with morphological operations to segment the overall breast volume from in-phase Dixon images. The method makes use of novel, problem-specific insights. The resulting segmentation mask was then applied to the corresponding Dixon water and fat images, which were combined to give Dixon MRI density values. Contemporaneously acquired T_1- and T_2-weighted image datasets were analysed using a novel and fully automated algorithm involving image filtering, landmark identification and explicit location of the pectoral muscle boundary. Within the region found, fat-water discrimination was performed using an Expectation Maximisation - Markov Random Field technique, yielding a second independent estimate of MRI density.

Results Images are presented for two individual women, demonstrating how the difficulty of the problem is highly subject-specific. Dice and Jaccard coefficients comparing the semiautomated BC-FCM method, operating on Dixon source data, with expert manual segmentation are presented. The corresponding results for the method based on T_1- and T_2-weighted data are slightly lower in the individual cases shown, but scatter plots and inter-class correlations for the cohort as a whole show that both methods do an excellent job in segmenting and classifying breast tissue.

Conclusions Epidemiological results demonstrate that both methods of automated segmentation are suitable for the chosen application and that it is important to consider a range of factors when choosing a segmentation algorithm, rather than focus narrowly on a single metric such as the Dice coefficient.

PACS numbers: 87.61.-c, 87.19.xj, 87.57.C-, 87.57.N-

Keywords: breast cancer, MRI, mammographic density, ALSPAC, segmentation

[a] Joint first author and corresponding author: Simon.Doran@icr.ac.uk

[b] Joint first author: j.hipwell@ucl.ac.uk

I. INTRODUCTION

Mammographic density, a quantitative measure of radio-dense fibrogladular tissue in the breast, is one of the strongest predictors of breast cancer risk. Women with more than 75% density have a four-fold or higher risk of breast cancer compared to those with less than 5%[1]. More intensive screening for women with high mammographic density has been proposed[2] but remains controversial[3].

However, in clinical practice, mammographic density, as assessed on x-ray mammograms, is generally reported using only qualitative, radiologist-assessed categories, and agreement between radiologists tends to be only moderate[4]. Quantitative analysis is hampered by the fact that breast density is an inherently 3-D material property and therefore not well suited to measurement using 2-D x-ray projections. Although subsequent risk assessment and epidemiological analysis rarely use full 3-D information (normally preferring a single number, i.e., the volume-averaged mean breast density), accurate derivation of such a statistic from the 2-D X-ray data is problematic and subject to error. Automated tools such as Volpara (VolparaSolutions, Wellington, NZ)[5] and QUANTRA (Hologic Inc., USA) are gaining traction in the mammography community, suggesting that mean breast density can be calculated without inter-reader bias. However, such readings may be affected by errors in estimating breast thickness[6] and the relation between the values of breast density reported and those obtained by other techniques remains to be elucidated[7].

Increasingly, Magnetic Resonance Imaging (MRI) mammography is being used in clinical and research settings to assess breast structure, because of its 3-D capabilities, its non-ionizing nature and the strong soft tissue contrast between fibroglandular (parenchymal) and fatty tissue. In an MRI context, breast density refers to the percentage of breast tissue volume that is deemed to be "parenchymal" and this is generally assumed to be the same as volume fraction of tissue whose MR signal arises from free water molecules, as opposed to fat (i.e., the "water fraction" or "percentage water"). Clearly, this is not an exact equivalent of the mammographic x-ray density. Nevertheless, Thompson *et al.*[8] demonstrate a clear correlation between the two.

At present manual evaluation of MRI 3-D breast density is an arduous, observer-dependent, and time-consuming process. Therefore, full or partial automation of the 3-D analysis of the breast is required. To achieve the desired segmentations of breast parenchymal volume and breast fat volume, two separate image processing tasks are required. First, the breast as a whole needs to be distinguished from the background and chest wall; and, second, the parenchymal tissue within the breast needs to be distinguished from fat.

Several different MRI pulse sequences have previously been used to assess breast density, but no definitive consensus has been reached about which is optimal. Few studies have compared different sequences within the same subject population. Furthermore, whilst there is a large body of prior literature (see Table I) describing different ways to achieve the two segmentation tasks described above, no studies, to date, have compared different automated methods with each other and with manual segmentation, for a sizeable subject population.

It is clear that many methods can produce "good" segmentation results. This study poses the following question: Do the minor differences we see between segmentations when we apply different algorithms on the same data actually matter for the uses to which the segmentations are ultimately put?

This study compares two very different methods of breast-outline segmentation: (i) an established[37] bias-corrected fuzzy C-means (BC-FCM) clustering technique based on a cost-function; and (ii) a new heuristic approach based on thresholding, landmark identification and direct analysis of image features. The results of this part of the study will be measures of overall breast volume from each method and volume similarity measures (Dice and Jaccard coefficients).

With the breast outline obtained, the second part of the study compares two methods of fat-water discrimination, again based on different principles: (i) The Dixon approach[38] uses scans acquired with an MRI technique that returns separate "fat" and "water" images. In principle, these allow us to obtain a fat and water fraction for every voxel, accounting for partial volume effects. However, Dixon sequences are not currently part of the routine acquisition protocol for clinical MRI examinations[39]. (ii) Our second method uses an analysis of the intensity histograms of the two different tissue classes in fat-suppressed T_1-weighted (T1w) and T_2-weighted (T2w) images. Such images are routinely acquired in diagnostic scanning and this method thus has the potential advantage of wider applicability if the two methods are shown to be concordant. Note that there is no means of obtaining ground truth data and, given that we are dealing with a healthy subject cohort, no possibility of obtaining x-ray data for comparison.

Nomenclature for the various segmentations is summarised in Figure 1.

TABLE I: Summary of journal papers describing methods to segment pectoral muscle and internal fibro-glandular tissue from MR images. N_{OB} refers to the number of observers who provided the gold standard manual segmentation. N_D indicates the number of MR data sets the method was validated with and N_S the number of MRI scanners. N/A = not applicable; N/S = not specified

Author, **Year**	**Ref. no.**	**Breast outline segmentation method**	**Fat / water classification method**	N_{Obs}	N_D	N_S
Hayton *et al.* 1997	9	Threshold, morphological opening followed by "dynamic programming"	None	N/S	3	N/S
Twellmann *et al.* 2005	10	Median filtering; Otsu automated thresholding; morphological closing	None	N/S	12	1
Koenig *et al.* 2005	11	Histogram-based threshold for breast-air, then Gaussian smoothing; intensity threshold for pectoral boundary, then min and max of locations with transition within confidence interval	None	N/S	4	N/S
Yao 2005	12	Threshold, morphological opening and region-growing followed by Bernstein-spline and active contour; automatic identification of key points to define rough surfaces of pectoral muscle; successive refinement via gradient-based technique, Bernstein spline and active contour	Fuzzy C-means	1	90	N/S
Lu *et al.* 2006	13	Region-growing, then spline and active contour for breast-air boundary; location of key points by geometry; identification of muscle slab, followed by spline	None	N/S	1	1
Giannini *et al.* 2010	14	Region-growing, then spline and active contour	None	2	12	2
Wang L *et al.* 2012	15	Hessian sheetness filter; 3-D connected component algorithm; intensity-based region-growing based on seed points automatically selected	None	1	84	5
Wu *et al.* 2012a,b, 2013a,b	16,17 18,19	Thresholding, morphological opening, contour extraction; three edge maps generated from original data and two non-linear filters; candidate selection; median filtering; dynamic time-warping; comparison between slices	Continuous Max-Flow	1	60	4

TABLE I: continued (atlas-based methods)

Author, **Year**	**Ref. no.**	**Breast outline segmentation method**	**Fat / water classification method**	N_{Obs}	N_D	N_S
Gubern–Mérida *et al.* 2011	[20]	Manually created atlas with 7 tissue classes; landmark detection	Bayesian atlas plus Markov Random Field regularisation	1	27	1
Gubern-Mérida *et al.* 2012, 2015	[21,22]	Manually created atlas; sternum detection; N3 bias-field correction	EM algorithm with Gaussian mixture model	3,4	27+23	1
Gallego-Ortiz and Martel 2012	[23]	Atlas created from Dixon in-phase images via entropy-based groupwise registration; maximal phase congruency and Laplacian mapping	None	N/S	500	1
Khalvati *et al.* 2015	[24]	Atlas created by manual initialization of active contour algorithm, subsequently corrected manually	None	N/S	400 + 17	3
Gallego and Martel 2011	[25]	Atlas, statistical shape model	None	N/S	415	N/S

TABLE I: continued (neural networks and fuzzy C-means)

Ertas *et al.* 2006, 2008	[26,27]	Breast air boundary: threshold; chest-wall: four cascaded cellular neural networks		1	39	N/S
Wang C-M *et al.* 2008	[28]	Support vector machines	Support vector machines	N/S	N/S	1
Wang Y *et al.* 2013	[29]	Support vector machines acting on multiple sets of MR images with different contrast	Support vector machines	N/S	4	1
Klifa *et al.* 2004, 2010	[30,31]	Fuzzy C-means	Fuzzy C-means	>1	30	N/S
Yang *et al.* 2009	[32]	Kalman filter-based linear mixing; fuzzy C-means	Kalman filter-based linear mixing	N/S	1	1
Nie *et al.* 2008	[33]	Fuzzy C-means; V-cut; skin-exclusion; B-spline; manual refinement via GUI	Fuzzy C-means	3	11	1
Sathya *et al.* 2012	[34]	Fuzzy C-means; support vector machines	None	N/S	1	1
Lin *et al.* 2011	[35]	Fuzzy C-means and B-spline fitting, building on[33], with inhomogeneity correction via N3	Fuzzy C-means, typically with 6 clusters	1	30	1
Lin *et al.* 2013	[33,36]	Template-based	As per[35]	1	30	1
Ertas *et al.* 2016	[37]	Bias-corrected FCM, followed by morphological opening and closing	None	1	82	¿4
This study		**Bias-corrected FCM vs thresholding, landmark analysis**	**Dixon vs T1w and T2w contrast**	**3**	**200**	**1**

A comprehensive epidemiological analysis of the relationship between breast composition and seven other physical, historical and lifestyle variables has been carried out for this cohort. Whilst the full report is beyond the scope of this study, we summarise the results and use them to discuss quantitatively the impact of differences between the various assessment methods on conducting reliable clinico-epidemiological studies.

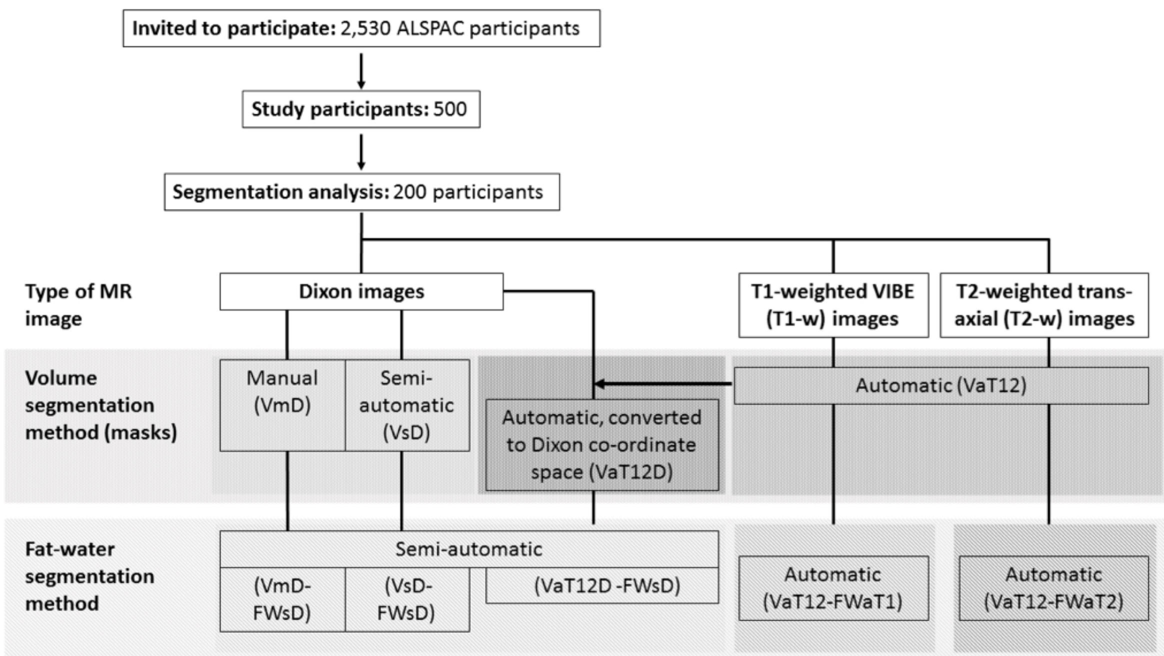


FIG. 1: Flow diagram of the overall data processing chain and nomenclature for the various segmentation methods. Some of these have the potential to operate on different source data and we can also combine the methods in different ways to achieve an overall result. We thus assign each step three codes: *segmentation purpose* (V = breast volume, FW = fat-water); *degree of automation* (m = manual, s = semi-automatic, a = fully automatic); and *source data* (D = Dixon; T1 = T_1-weighted, T2 = T_2-weighted, T12 = uses both T_1- and T_2-weighted data). Thus, a breast-volume measurement using semi-automatic segmentation on original Dixon data would be represented as VsD. Fat-water segmentations require both source data and a previously-generated volume mask, so are represented by the combination of two codes. For instance, fat-water statistics calculated semi-automatically from Dixon source data and using a mask generated automatically from T1w and T2w data would be described by VaT12-FWsD. We note one additional case, in which the volume mask VaT12 is re-sampled to give a result in the same coordinate space as the Dixon images and we assign this the label VaT12D.

II. METHODS

A. Data

1. Study Population

This work forms part of an investigation into breast composition at young ages, nested within the Avon Longitudinal Study of Parents and Children (ALSPAC). ALSPAC originally recruited 14,541 pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992, as described by Boyd *et al.*in a cohort profile paper[40]. For this sub-study, Caucasian nulliparous women were invited to attend an MRI examination at the University of Bristol Clinical Research and Imaging Centre (CRIC) between June 2011 and November 2014. Women were restricted to those from a singleton birth, who had never been diagnosed with a hormone-related disease and had regularly participated in follow-up surveys, including completing the age 20y questionnaire (2010-2011). Of the 2530 invited, 500 (19.8%) eligible women attended.

The ALSPAC Law and Ethics Committee and the Local Research Ethics Committees gave ethical approval for the study. The study website contains details of all the data that are available through a fully searchable data dictionary[41].

2. MR Imaging

Participants underwent a breast MRI scan using a 3T Siemens Skyra MR system with a breast coil that surrounds both breasts of a prone patient. Three sets of bilateral images were acquired:

- multislice, sagittal Dixon[38] images (in-phase, out-of-phase, water and fat), acquired using a turbo spin-echo sequence with nominal in-plane resolution of (0.742×0.742) mm^2, nominal slice thickness 7 mm and interslice spacing 7.7 mm;
- T1-weighted 3D images, acquired using a VIBE sequence with fat saturation and a nominal resolution of $(0.759 \times 0.759 \times 0.900)$ mm^3, as routinely used in clinical dynamic contrast-enhanced MRI protocols for the breast;
- multislice, axial, T2-weighted images, acquired using a turbo spin-echo sequence, with

nominal in-plane resolution of (0.848×0.848) mm^2, and both slice thickness and spacing between slices 4 mm;

3. Manual Reference Segmentation

To assess breast volume, a manual segmentation protocol (as described in the Supplementary Information)was developed and used by three readers (RD, MB and ISS) independently to outline the breast from surrounding tissues in the Dixon images, using ITK-SNAP (version 3.0.0). All subjects had a manual segmentation of all breast slices performed by at least one reader. The datasets of 16 representative subjects were manually segmented twice by all three readers to assess between- and within-observer variation. In cases where more than one manual segmentation is performed, the VmD and VmD-FWsD results quoted below represent the median values taken for the multiple manual readings.

4. Training and Validation Data Sets

A training set of 100 randomly selected subjects was used to make initial comparisons across MR images and segmentation methods, and for the manual readings, between- and within-observer variation. The training data were used to assess the common reasons for segmentation failure and to improve the algorithms. At the end of the testing phase, the algorithm code was "frozen" and final comparisons of the segmentation methods were completed on a second set of images from a further 100 participants. Except where stated otherwise, all the summary statistical results presented here come from this second, "validation" cohort. For further details concerning statistical methods, please see the Supplementary Information.

B. Breast Outline Segmentation

1. Semi-automated, bias-corrected fuzzy C-means (BC-FCM)

A fuzzy C-means (FCM) algorithm was applied to the Dixon in-phase images. It has the advantage that it can be modified to carry out a simultaneous intensity inhomogeneity compensation, or bias-correction (BC), and this is potentially less expensive computationally than a prefiltering operation[42]. The algorithms in this section were implemented using IDL (Harris Geospatial Systems, Melbourne, FL, USA) and run on a standard desktop computer.

The BC-FCM variant we implemented is described in[37]. Formally, the algorithm does not require a training dataset and so is an unsupervised clustering algorithm. However, in practice, some experience with the types of data involved can improve the results dramatically. Except for the local smoothness criterion (introduced by cost function γ in ref.[37] — see this publication for all other related notation), BC-FCM *per se* does not use any spatial information. Nevertheless, a "good" segmentation involves a number of problem-specific insights and the basic BC-FCM method above was enhanced by additional heuristic algorithms in the spatial domain, based on the results obtained with the training data.

a. Initial parameters and iteration threshold After some experimentation, $\beta(\mathbf{r})$ was set to 0.1 for all spatial locations and ϵ to 0.01. The two initial class centroids c_f were calculated by taking the mean of the slice being processed and adding a lower and an upper offset. These two offsets are adjustable parameters under user control. For many subjects — see the Results section for an example —, a single set of defaults performed extremely well. However, for a small subset of "difficult" cases — second example in Results —, user interaction was needed to try various combinations. As implemented here, on a standard desktop computer, running non-optimised software, it took around 2 mins. to run the segmentation algorithm on each 3-D dataset. Thus, this "trial and error" step was the most frustrating feature of the BC-FCM method in practice. Numerous coding and hardware improvements (e.g., parallelisation) could be made to the prototype to improve the user experience, potentially allowing these adjustable parameters to be altered by simple slider controls with immediate feedback.

We observed an improvement in performance by allowing the algorithm to perform separate BC-FCM classifications for segmenting the posterior of the breast from the chest wall and segmenting the anterior portion from air, then merging the two volumes. Furthermore, it was noted that the optimal offsets providing the initial class centroids were often different for these two segmentation problems. Thus, each dataset is split into two portions in an anterior-posterior (AP) direction and the BC-FCM algorithm applied twice per image slice. Given that the size of breasts varies, the position of the AP-split is also different for different datasets and this is handled automatically by having two passes through the entire algorithm with an automated choice of the AP-split position made after Pass 1.

b. Morphological operations The breast outlining task requires a definite boundary to be drawn. Thus, it is not necessary to use the full membership function output of the BC-FCM routine, and we arrange for the clustering to produce a binary image. This may include some misclassified regions outside the breast and some "holes" inside the breast. To remove the unwanted regions, 2D hole-filling followed by a 4-neighbourhood connectivity search and object labelling is performed. The largest non-background object in each slice is identified as the breast region and other smaller objects are removed from the binary image. This exercise is repeated for all slices and these are then merged to form an approximate breast volume.

Within this approximate breast volume, there may be some non-breast tissue segmented for cases in which fatty breast tissue is connected to the chest and liver; and there may also be some unsegmented breast tissue left for cases in which dense breast tissue is connected to the chest wall muscles. To reduce these over- and under-segmentations, 3D morphological image opening is performed, followed by closing using two cylindrical structuring elements having the same radius of 3 voxels but different heights of 3 voxels and 25 voxels in the axial direction. These parameters were found by experimentation during our previous study[37].

c. Lateral cutoffs The preceding steps in the process do an excellent job in segmenting the anterior and posterior margins of the breast. However, there is no consensus in the literature as to "where the breast stops" in the right-left and superior-inferior directions. The extent of the breast is not directly delineated by any change in MRI contrast and the required boundary may, indeed, be specific to the application of the imaging (e.g., when comparing the MRI segmentation with the breast region compressed within the paddles of a mammography system, the axilla region may be excluded entirely). Thus, based on the consensus protocol (Appendix **??**) reached by the three experienced readers, a heuristic algorithm was developed, as described below. This additional truncation is derived entirely from geometric considerations and boundaries are drawn without regard to image intensity, which is in many cases the same on either side of the boundary.

Each breast is processed in turn. The stack of sagittal images segmented using BC-FCM forms a pseudo 3-D dataset. From this dataset the transverse plane containing the largest breast area is passed to a simple algorithm that extracts the air-breast interface as a 1-D "breast profile". (This geometry is illustrated as Figure S2 of the Supplementary Information.)The profile is used to determine the position of the breast midpoint in a left-right direction. Working outwards from this midpoint, we find the first position at which the absolute value of the gradient (approximated by the finite difference between adjacent voxels) of the breast profile rises above a threshold value, determined by experimentation. This indicates a change in angle of the skin surface from flat regions between and outside the breasts, to the side contour of the breast. A mask is applied to exclude all sagittal slices in the original dataset on either side of these changes in angle. (Typically, the "raw" output of the BC-FCM algorithm would include these.) Finally, a similar profile is generated for the superior-inferior direction and the upper and lower bounds of the breast are determined in each sagittal plane of the original data.

2. Fully-automated, using T1w and T2w Images

a. Pre-Processing Processing (Bias-Field Correction) A slowly varying bias-field, caused by inhomogeneities in the magnetic field during the MR acquisition, is a common artefact of MR images. To correct this for the T1w and T2w images, we apply the "N4ITK" nonparametric non-uniform intensity normalization method[43]. This is a refinement of the popular N3 algorithm which adopts a fast, robust B-spline fitting algorithm and a hierarchical, multi-scale, optimisation scheme (figures 2a and 2b).

b. Breast Mask Segmentation This novel, heuristic method, implemented using the Insight Toolkit[44], computes a whole breast mask using both the T1w and T2w images. In developing this automated approach, emphasis has been placed on limiting the number of empirically derived parameters and relying instead on detecting statistical or functional extrema. In this way we aim to make the method as widely applicable to variations in subjects and images as possble. The method comprises a number of distinct processing steps as follows.

1. The T2w image is resampled to match the resolution of the T1w image.

2. A grey-scale closing operation along each of the orthogonal axes, $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$, is performed on the T2w image, to eliminate voids from the subsequent foreground segmentation. In this operation each voxel's intensity, I_{T2w}, at index (i, j, k) is replaced by $I_{\mathrm{cT2w}}(i, j, k)$ according to:

$$I_{\mathrm{cT2w}}(i,j,k) = \min\Bigg[\min\left(\max_{0\le i_1\le i} I_{\mathrm{T2w}}(i_1,j,k),\ \max_{i<i_2<N_i} I_{\mathrm{T2w}}(i_2,j,k)\right), \\ \min\left(\max_{0\le j_1\le j} I_{\mathrm{T2w}}(i,j_1,k),\ \max_{j<j_2<N_j} I_{\mathrm{T2w}}(i,j_2,k)\right), \\ \min\left(\max_{0\le k_1\le k} I_{\mathrm{T2w}}(i,j,k_1),\ \max_{k<k_2<N_k} I_{\mathrm{T2w}}(i,j,k_2)\right)\Bigg] \tag{1}$$

where N_i, N_j, N_k are the number of voxels along each axis.

3. The T1w image is rescaled to match the intensity range of the closed T2w image and the maximum of these two images, $I_{\mathrm{MaxT1wT2w}}$, computed.

4. The foreground (i.e. the subject) is segmented from the background by thresholding, $I_{\mathrm{MaxT1wT2w}}$. The threshold, t_{bg}, is computed via:

$$t_{\mathrm{bg}} = \arg\max_{I} \left[F_{\mathrm{dark}}(I)\left(F_{\mathrm{CDT}}(I) - F_{\mathrm{var}}(I)\right)\right] \tag{2}$$

according to the following functional criteria:

   - The background is assumed dark therefore the threshold should be close to zero:

$$F_{\mathrm{dark}}(I) = 1 - \frac{I}{\max(I)} \tag{3}$$

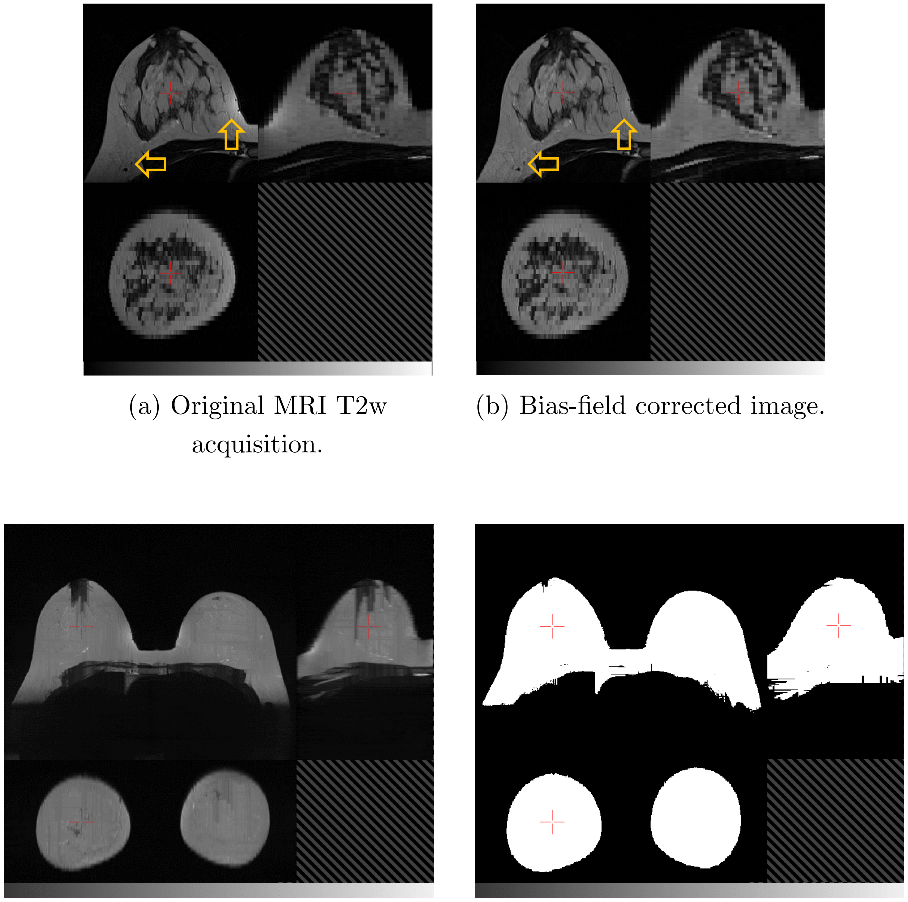
   - The frequency of voxel intensities in the background is higher than the foreground i.e. the background intensities form a distinctive peak in the image histogram, P(I), which is captured by a sharp rise in the cumulative intensity distribution function:

$$F_{\mathrm{CDT}}(I) = \frac{\sum_{j=0}^{I} P(j)}{\sum_{k=0}^{\max(I)} P(k)} \tag{4}$$

   - The background has a lower intensity variance than the foreground:

$$F_{\mathrm{var}}(I) = \frac{\sum_{j=0}^{I} P(j)(j-\mu)^2}{\sum_{k=0}^{\max(I)} P(k)(k-\mu)^2} \tag{5}$$

The resulting foreground mask image is denoted I_{fg} — see Figure 2(d).



(a) Original MRI T2w acquisition.

(b) Bias-field corrected image.

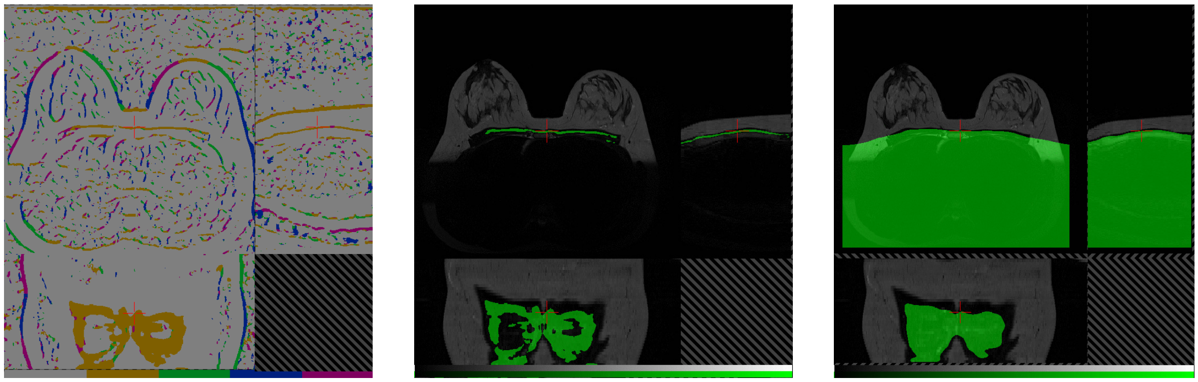
(c) Closed image.

(d) Foreground mask.

FIG. 2: Orthogonal slices through (a) a T2 weighted MRI and (b) the corresponding image after bias-field correction, with arrows indicating regions that are particularly improved by the processing. The "closed" T2w image is shown in (c) and foreground mask I_{fg} in (d). In each image the top-left quadrant is the axial slice, the top-right is sagittal and the bottom-left is coronal.

5. Landmark identification. The most anterior voxels in the foreground mask, I_{fg}, on the left and right sides of the volume, are identified and assumed to be approximately coincident with the nipple locations. If multiple voxels are found then the center of mass of the cluster is computed. The mid-sternum is computed as the most anterior voxel of the foreground mask, equidistant from the nipple landmarks in the coronal plane.

6. Pectoral muscle boundary extraction. Various methods have been presented in the literature to segment breast MRI volumes and the pectoral muscle (Table I). These include semi-automated methods requiring user interaction[31,33,36], 2D mid-slice template registration[36], statistical shape models[25] and atlas-based methods[16,18–20,24,45].

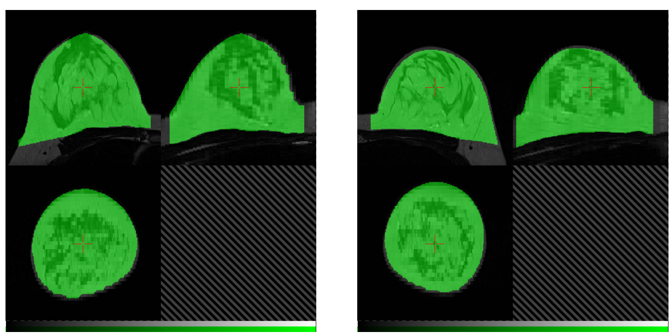


(a) Detected “dark line” structures.

(b) Pectoral mask.

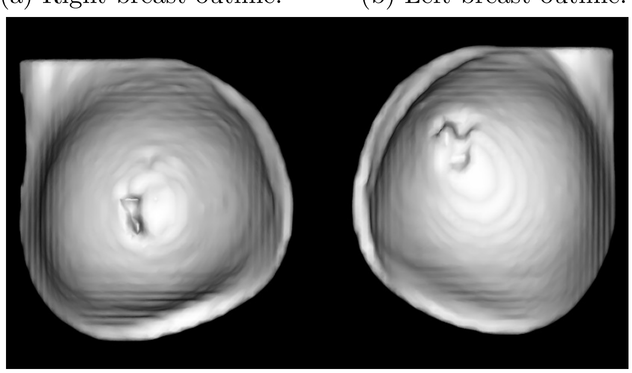
(c) Extrapolated B-Spline surface mask.

FIG. 3: The anterior pectoral muscle surface is detected using the Oriented Basic Image Feature “dark line” class. Subplot (a) shows these features detected at four orientations (OBIF_{15} to OBIF_{18}). Region-growing the “brown” medial-lateral class, OBIF_{15}, closely delineates this anterior boundary immediately posterior to the sternum (b). The anterior surface of this mask is extrapolated using a B-Spline fit to the lateral boundaries of the volume (c).



(a) Right breast outline.

(b) Left breast outline.



(c) Surface rendering.

FIG. 4: Breast region mask created by removing the pectoral surface mask (figure 3c) from the foreground mask (figure 2d). Two views of the mask are shown, superimposed on the original MR image and centered on the right (a) and left (b) breasts. The surface rendering (c) illustrates the “squaring off” to include the axilla.

A number of methods have been developed to segment explicitly the pectoral muscle. These include a B-spline fit to the intensity gradient of the pectoral boundary[33], anisotropic diffusion and Canny edge detection[17] and Hessian matrix planar shape filtering[15,46]. Atlas-based methods have been shown to perform well but are computationally intensive[47] and require significant initial investment of time to develop a library of atlases.

We have developed a method to detect explicitly the anterior pectoral muscle boundary in individual MR volumes. Our approach has similarities to the Hessian processing of Wang *et al.*[15,46], in that it employs Gaussian derivatives to detect regions in the image with a planar profile. However rather than computing a ratio of the eigenvalues of the Hessian matrix and thresholding the result, we obtain a direct classification of linear structures, immediately posterior to the sternum, using Oriented Basic Image Features (OBIFs, Figure 3).

The concept of Basic Image Features (BIFs) was developed by Griffin [48]. The technique classifies pixels in a 2D image into one of seven classes according to the local zero-, first- or second-order structure. This structure is computed using a bank of six derivative of Gaussian filters (L_{00}, L_{10}, L_{01}, L_{20}, L_{11} and L_{02}) which calculate the nth (where n=0,1,2) order derivatives of the image in x and y (S_{00}, S_{10}, S_{01}, S_{20}, S_{11} and S_{02}). By combining the outputs of these filters, any given pixel can be classified according to the largest component of vector BIF:

$$\mathrm{BIF} = \left\{ \begin{array}{l} \overset{\text{flat}}{\epsilon S_{00}}, \overset{\text{slope−like}}{2\sqrt{S_{10}^2 + S_{01}^2}}, \\ \overset{\text{maximum}}{\lambda}, \overset{\text{minimum}}{-\lambda}, \\ \overset{\text{light line}}{\frac{\lambda+\gamma}{\sqrt{2}}}, \overset{\text{dark line}}{\frac{\lambda-\gamma}{\sqrt{2}}}, \overset{\text{saddle}}{\gamma} \end{array} \right\} \tag{6}$$

given

$$\lambda = \sigma^2 \frac{(S_{20} + S_{02})}{2} \tag{7}$$

$$\gamma = \sigma^2 \sqrt{(S_{20} + S_{02})^2 + 4S_{11}^2} \tag{8}$$

$$\tag{9}$$

In addition, slopes, light lines, dark lines and saddles can be characterised according to their orientation (OBIFs). We quantise this orientation into four, 45 degree quadrants which produces eight slope sub-classes (OBIF_1 to OBIF_8), and four sub-classes for each of light lines (OBIF_{11} to OBIF_{14}), dark lines (OBIF_{15} to OBIF_{18}) and saddles (OBIF_{19} to OBIF_{22}).

By region-growing the medial-lateral, OBIF_{15} dark line features detected in each axial image slice, in 3-D, from seed positions immediately posterior to the mid-sternum, we obtain a binary segmentation of the anterior pectoral muscle surface. The BIF processing was performed at a single scale using a Gaussian kernel with standard deviation 5 mm. A smooth B-spline surface is then fitted to the anterior voxels of the resulting mask[44] to extrapolate the muscle surface to the lateral boundaries of the image volume (figure 3c).

7. Finally we generate a 2D coronal mask, I_{CNL}, to crop non-breast tissue from the whole breast mask. I_{CNL} is computed from a coronal skin elevation map, I_{skin2D}, which contains the distance of each anterior skin voxel in the foreground mask, I_{fg}, from the most posterior boundary of the MR volume. The coronal profile of each breast is obtained by thresholding I_{skin2D} at

$$h = \frac{(4h_{\mathrm{ms}} + h_{\mathrm{Ln}} + h_{\mathrm{Rn}}}{6} \tag{10}$$

where h_{ms} is the anterior elevation of the mid-sternum landmark, and h_{Ln} and h_{Rn} are the left and right nipple anterior elevations respectively. The roughly circular profile obtained for each breast is then dilated by 10mm and the mask squared off, to create a superior-lateral corner and hence extend the breast volume into the axilla (figure 4c)

C. Fat-Water Discrimination

1. *Semi-automated calculation of percentage breast density, based on Dixon Images*

In principle, the output from a Dixon pulse sequence is a set of images reflecting water content $I_w(\mathbf{r})$, which we identify with the parenchymal component of the breast, and an equivalent set $I_f(\mathbf{r})$ reflecting fat content. Ideally, these images would be quantitative and allow the direct calculation of the water and fat fractions $\phi_w(\mathbf{r})$ and $\phi_f(\mathbf{r})$ via the equation[49]

$$\phi_w = \frac{I_w}{I_w + I_f} \qquad \text{and} \qquad \phi_f = \frac{I_f}{I_w + I_f} \tag{11}$$

In practice, there are a number of complicating factors:

- Parenchymal tissue and fat have different relaxation properties and, since the acquisitions are not generally designed to be proton density weighted, this means that the relative intensities of equal fractions of fat and water are different.
- The B_1 field of the probe is not uniform across the whole breast and this leads to a spatially-dependent efficacy of the fat-water separation.
- In practice, the fat tissue does not have a single proton resonance.
- Different manufacturers have different proprietary image reconstruction methods and these may influence the quantitative results.

Our solution to (at least) the first of these problems is to proceed as follows:

(a) Identify a small region in the water image that is expected to be entirely composed of parenchymal tissue. The region should be in a part of the image that is free from intensity artefacts caused by proximity to the RF coil (i.e., the data should come from a homogenous region of B_1).

(b) In the fat image, identify similarly a second region entirely composed of fat.

(c) Calculate the ratio of the average voxel values in each of the two regions:

$$r \;=\; \frac{1}{N_w} \sum_{i \in \mathrm{ROI}_w} I_w(\mathbf{r}_i) \Big/ \frac{1}{N_f} \sum_{j \in \mathrm{ROI}_f} I_f(\mathbf{r}_j) \tag{12}$$

where N_w and N_f are the numbers of voxels in the selected regions-of-interest ROI_w and ROI_f respectively.

(d) Replace the value I_f in Eq. (11) with rI_f.

This procedure potentially improves the accuracy of the water-fraction calculation but at the cost of introducing an interactive step into the density estimation process. We have not tested in a systematic fashion the influence that the size and shape of the region-of-interest selection have on the process, in part because we have no ground truth values. A further issue with this technique is that in the limiting cases of extremely dense or extremely fatty tissues, it may not be possible to find appropriately "pure" regions of both types.

2. Fully-automated, using T1w and T2w Images

Fuzzy c-means (FCM) clustering has been evaluated by a number of studies to classify the internal structure of the breast into fat and fibro-glandular tissue classes[16,18,29,31,33–35,50] Table I). Song *et al.*[50] adopt a Gaussian kernel FCM, whilst Sathya[34] use a quadratic kernel FCM to train a support vector machine (SVM). In[29], Wang *et al.* use a multi-parametric hierarchical SVM classification approach to segment the internal breast and found this to be superior to both a conventional SVM[28] and FCM segmentation. T1W, T2W, proton density and three point Dixon (water and fat) images were all incorporated. Klifa *et al.*[31] compared the resulting volumentric MRI density measurement of their method with mammography but found only modest correlation ($R^2 = 0.67$).

In[20] a probabilistic atlas approach was proposed. This requires a sizeable number of pre-labelled atlases to be created, considerable computation to register them and assumes correspondence between fibro-glandular structures across the population. To address the latter a Markov Random Field (MRF) was introduced to spatially regularise the classification of each voxel according to that of its neighbours. Similarly Wu *et al.*[16] use the registered atlas as a pixel-wise fibroglandular likelihood prior for a multivariate Gaussian mixture model and demonstrate superior performance when compared to FCM using a manual thresholding approach as the gold standard. In a later publication[19], the same authors investigate a continuous max-flow (CMF) algorithm to generate a voxel-wise likelihood map using the same atlas initialisation. They demonstrate that this approach performs better with the atlas initialisation than without, but that FCM is superior to the CMF approach without the atlas.

Mixture models have also been proposed by Yang *et al.*[32] who implement a method using Kalman filter-based linear mixing. They demonstrate it out-performs a c-means method but evaluation using real MR data was limited.

Our segmentation of the T1 and T2 MRI data into fat and glandular tissue is a modification of that proposed by Van Leemput *et al.*[51] in which an intensity model and spatial regularization scheme are optimized using a Maximum Likelihood formulation of the Expectation-Maximisation (EM) algorithm. The EM algorithm iteratively updates the Gaussian probability distributions used to estimate the intensity histograms of each tissue class (fat and non-fat) via a Maximum Likelihood formulation. In order to improve classification of voxels in which the partial volume of fat and glandular tissues is a significant factor, a Markov Random Field (MRF) regularization scheme is employed to ensure spatial consistency. The MRF modifies the probability of a particular voxel being assigned to either the fat or glandular classes (or a proportion of either) according to the current classification of neighbouring voxels. In this way isolated regions of glandular tissue in very fatty regions, for instance, are penalized in favour of a more realistic and anatomically correct arrangement of the classes.

D. Epidemiology

Appropriate linear and logistic regression models were used to examine associations of average total breast, fat and water volumes, and percent water, as measured using different MR images and segmentation methods, with selected established and potential mammographic density correlates. Breast measures were log-transformed and the exponentiated estimated regression parameters represent the relative change (RC) in breast measure with a unit increase, or category change, in the exposure of interest (with 95% confidence intervals (95% CI) calculated by exponentiating the original 95% CIs). Age at menarche (months), height (cm) and BMI (height (cm)/ weight (kg)^2) at MR were treated as continuous variables and centred at the mean. Current hormone contraceptive use, cigarette smoking and alcohol drinking were treated as binary (yes/no) variables. Mothers mammographic density (%) was averaged between both breasts, and maternal age (months) at mammography and clinically measured or self-reported maternal BMI (median 3 years (inter-quartile range (IQR) = 1.5 years) prior to mammography)) were used as continuous measures and centred at the mean. Variables were included as potential determinants of breast measures, or as confounding factors, where appropriate.

Data analysis was conducted with STATA statistical software, Version 14.

III. RESULTS

A. Breast Outline Segmentation

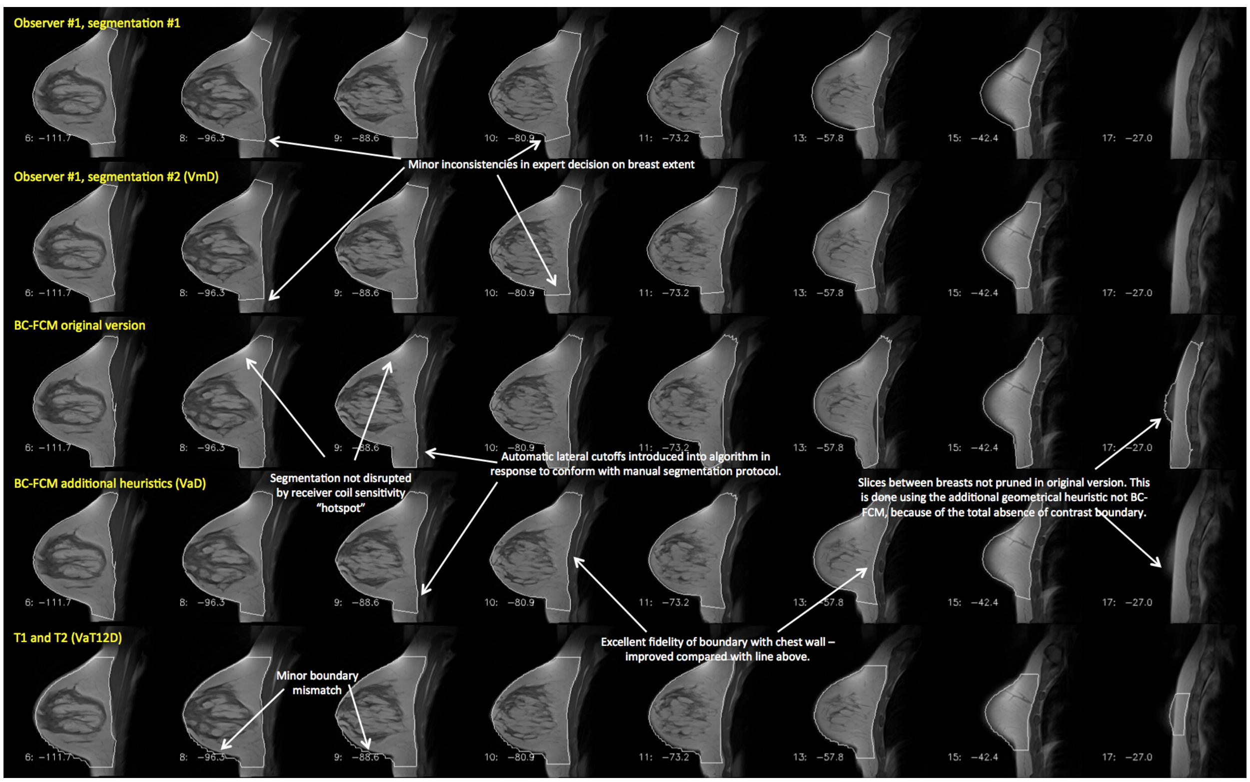


FIG. 5: : Example of a case where both of the algorithms examined in this work performed well. Features of interest in the various different segmentations are annotated. *Note that this image is provided with high resolution and can be zoomed significantly to reveal additional detail.*

Figure 5 shows an example of the two methods applied to a dataset containing medium-sized breasts, with a moderate parenchymal content. There is a border of fat around the parenchyma, which, at the posterior of the breast, leads to excellent contrast at the boundary with the chest wall, making segmentation a relatively straightforward task. Results are

TABLE II: : Dice and Jaccard coefficients for the "easy" segmentation problem of Fig. 5. Note that the BC-FCM/heuristics (VaD) represents the fully automated version, running with default parameters.

	Manual 1	Manual 2	BC-FCM Orig	BC-FCM /heuristics (VaD)	VaT12D
Dice Coefficients					
Manual 1	1.000				
Manual 2	0.949	1.000			
BC-FCM Orig	0.854	0.877	1.000		
BC-FCM/heuristics (VaD)	0.901	0.924	0.921	1.000	
VaT12D	0.887	0.888	0.810	0.865	1.000
Jaccard Coefficients					
Manual 1	1.000				
Manual 2	0.904	1.000			
BC-FCM Orig	0.745	0.781	1.000		
BC-FCM/heuristics	0.820	0.859	0.853	1.000	
VaT12D	0.797	0.799	0.681	0.761	1.000

shown for two separate manual segmentations by the same experienced observer; for the BC-FCM method from ref.[37]; the BC-FCM method with additional heuristics and default parameters, as described above; and the new method based on T1 and T2 images (VaT12). It will be seen that the segmentation performance is excellent, with only minor difference between the methods. Note how implementation of guidelines developed during the manual segmentation process supplements the BC-FCM approach in order to cut off the segmentation in both the left-right and superior-inferior directions, where there are no corresponding intensity boundaries seen in the image data themselves.

Table II shows the Dice and Jaccard coefficients for the four sets of segmentations illustrated in Figure 5, confirming the excellent performance of all the algorithms.

By contrast, Figure 6 illustrates a case where all assessment methods have far more difficulty in providing a correct segmentation. Smaller breasts tend to be more problematic to segment, as a higher fraction of the segmentation involves partial-volume effects. Highly parenchymal breasts have very low (sometimes no) contrast between the parenchyma and pectoral muscles of the chest wall, and the intensity-based BC-FCM algorithm has particular difficulties in this regard. Many slices require a high degree of anatomical knowledge to perform the segmentation. Consider the two versions of the BC-FCM results presented. With the default parameters in the upper of the two rows, over-segmentation occurs in slice

TABLE III: : Dice and Jaccard coefficients for the difficult segmentation problem of Fig. 6

	Manual 1	Manual 2	BC-FCM Orig	BC-FCM /heuristics (best)	BC-FCM Edited (VsD)	VaT12D
Dice Coefficients						
Manual 1	1.000					
Manual 2	0.915	1.000				
BC-FCM Orig	0.776	0.797	1.000			
BC-FCM /heuristics (best)	0.836	0.792	0.782	1.000		
BC-FCM Edited (VsD)	0.914	0.913	0.809	0.828	1.000	
VaT12D	0.796	0.771	0.728	0.818	0.795	1.000
Jaccard Coefficients						
Manual 1	1.000					
Manual 2	0.843	1.000				
BC-FCM Orig	0.634	0.662	1.000			
BC-FCM /heuristics (best)	0.718	0.657	0.642	1.000		
BC-FCM Edited (VsD)	0.842	0.840	0.679	0.707	1.000	
VaT12D	0.661	0.627	0.572	0.692	0.660	1.000

11 and part of the chest wall is included in the parenchymal breast region. By contrast, with the “best” set of parameters (as found by repeating the algorithm and manually adjusting them), the lower row shows that the problem in slice 11 is corrected, with good matching of the pectoral muscle contour, but only at the cost of introducing an under-segmentation in slice 8, and, worse, losing the segmented breast region entirely in slice 6. In practice, where such problems occurred, it was necessary to edit the final segmentations manually. (Note on terminology: As shown in Fig. 6, the “BC-FCM/heuristics (VaD)” method cannot reliably be run for the whole cohort using only default parameters and so we must describe the technique as semi- rather than fully-automated. Even for cases where no manual editing or parameter adjustment need to be performed, human inspection is still required to confirm this. All subsequent cohort statistics will therefore use the nomenclature VsD to reflect this.)

We have run a similar analysis on all 16 cases for which we have duplicate manual segmentations by all three observers. The detailed results are shown in the Supplementary Information.

A second method of examining the relation between the volume segmentation results is to plot the total breast volume obtained by one method against that of another. In the

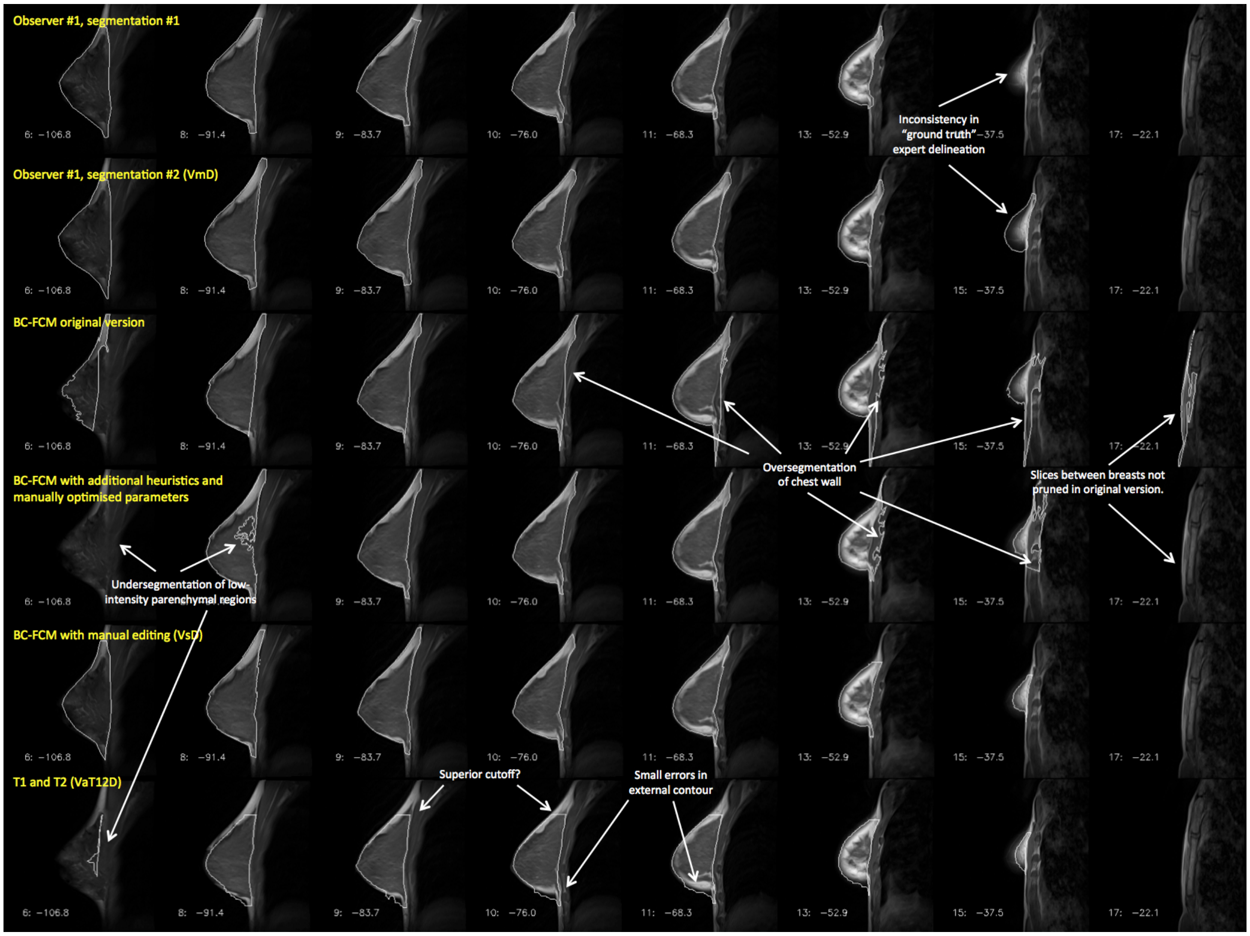
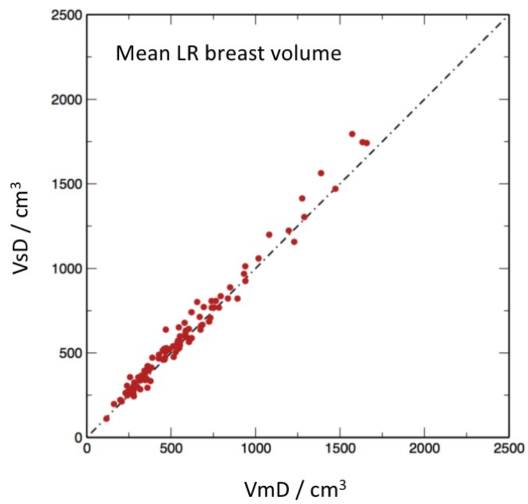
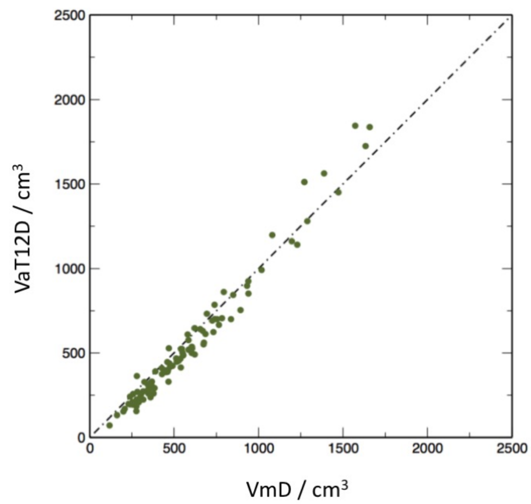


FIG. 6: : Example of a case where automatic segmentation is difficult. The rows represent the results of different segmentations and, for compactness, an informative subset of slices has been chosen to illustrate important features of the problem. *Note that this image is provided with high resolution and can be zoomed significantly to reveal additional detail.*

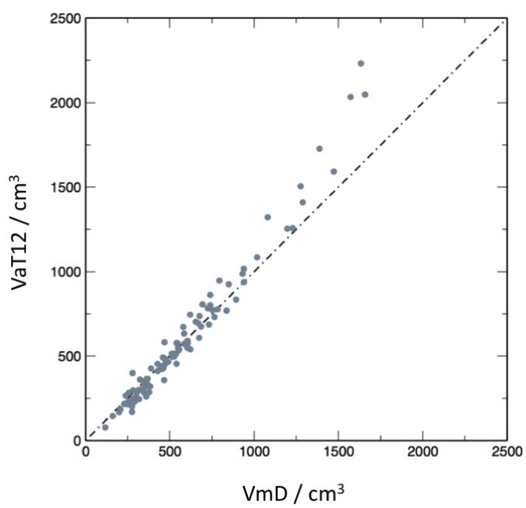
scatter plots of Figures 7(a)–(c), the x- and y-coordinates of each point represent the mean, for a single subject, of the left and right breast volumes evaluated, respectively, by the two methods under consideration. Figure 7(a) compares VsD, the semi-automated BC-FCM method using Dixon image input, with the "gold-standard" median manual segmentation, VmD, measured on the same Dixon dataset. Figure 7(b) gives results for the VaT12 method, which operates on the T1w and T2w datasets and evaluates the breast volume in the coordinate space of the T1w dataset. Finally, Figure 7(c) looks at the effect of resampling the map generated by the algorithm in (b) with the spatial resolution and frame of reference of the Dixon data, which we term VaT12D. In each case, the line of identity is shown and Table IV reports the corresponding inter-class correlations (ICC), representing the proportion of variance across participants shared between different ascertainment methods.



(a)



(b)



(c)

FIG. 7: Scatter plots of mean left and right breast volumes in cm^3 for the different methods in comparison to manual segmentation: (a) volume from semiautomatic segmentation of Dixon images (VsD) vs volume from manual segmentation (VmD); (b) volume via automated segmentation from T_1- and T_2-weighted images transformed to Dixon reference frame (VaT12FD) vs manual (VmD); (c) volume obtained from T_1- and T_2-weighted images in native 3-D reference frame (VaT12).

TABLE IV: : Inter-class correlations for total breast volume segmentations.

	VmD	VsD	VaT12D	VaT12
VmD	1.000			
VsD	0.990	1.000		
VaT12D	0.974	0.977	1.000	
VaT12	0.985	0.992	0.982	1.000

TABLE V: : Inter-class correlations for total water volume segmentations.

	VmD-FWsD	VsD-FWsD	VaT12D-FWsD	VaT12-FWaT1	VaT12-FWaT2
VmD-FWsD	1.000				
VsD-FWsD	0.995	1.000			
VaT12D-FWsD	0.992	0.993	1.000		
VaT12-FWaT1	0.920	0.921	0.924	1.000	
VaT12-FWaT2	0.948	0.949	0.962	0.899	1.000

B. Fat-Water Segmentation

Figures 8 and 9 present the results of the fat and water segmentation in the same format as for the total breast volume. In this case, however, a further option is available. Although the breast outline segmentation VaT12 requires both the T1w and T2w data, once this mask is available, it is possible to obtain two separate fat-water segmentations one using just the T1w and one using just the T2w data. These are denoted VaT12-FWaT1 and VaT12-FWaT2 respectively.

The inter-class correlation (ICC) for total water volume, representing the proportion of variance across participants shared between the different ascertainment methods, are given in table V.

C. Epidemiological Results

A diagrammatic summary of the results of the epidemiological analysis is presented in Figure 10 and further details of the work are reported as supplementary information.

Associations with both breast volume and breast water fraction were found for current body mass index (BMI). For a 1 kg m^{-2} increase in BMI, a relative change in breast volume of 1.13[1.10, 1.16] was observed for the cohort for both the VmD and VsD methods and the corresponding result for the VaT12 family of methods was 1.15[1.12, 1.18], where the figures in square brackets are the 95% confidence intervals. A smaller, but still important, decrease in breast water fraction was seen, and the corresponding statistics are VmD-FWsD, VsD-FWsD 0.96[0.95, 0.97], VaT12D-FWsD 0.95[0.94, 0.97], VaT12-FWaT1 0.97[096, 098], VaT12-FWT2 0.95[0.94, 0.96].

A weak association between current height and breast volume was also observed. For a

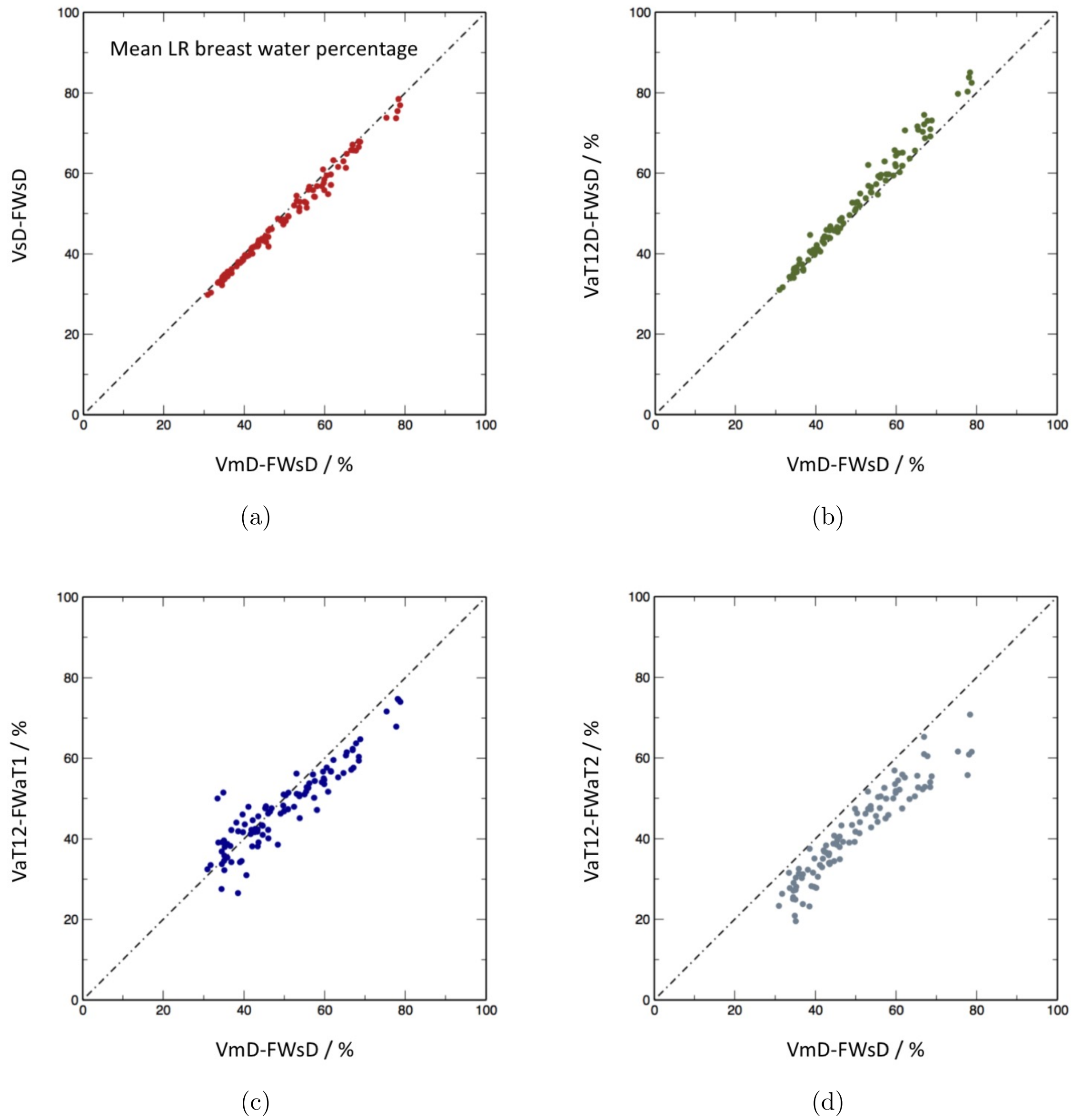


FIG. 8: Scatter plots of mean left and right breast water percentage for the different methods in comparison with manual segmentation on Dixon images followed by percentage water estimation the using semiautomated Dixon image method: (a) semiautomatic segmentation of Dixon images followed by percentage estimate from Dixon image data (VsD-FWsd); (b) volume via automated segmentation from T_1- and T_2-weighted images transformed to Dixon reference frame (VaT12FD) followed by semiautomated percentage estimate from the Dixon data (VaT12D-FWsd); (c) volume obtained from T_1- and T_2-weighted images in native 3-D reference frame, followed by automatic percentage estimate from T_1-weighted data (VaT12-FWaT1); (d) as (c), but with the water percentage estimated from the T_2-weighted data.

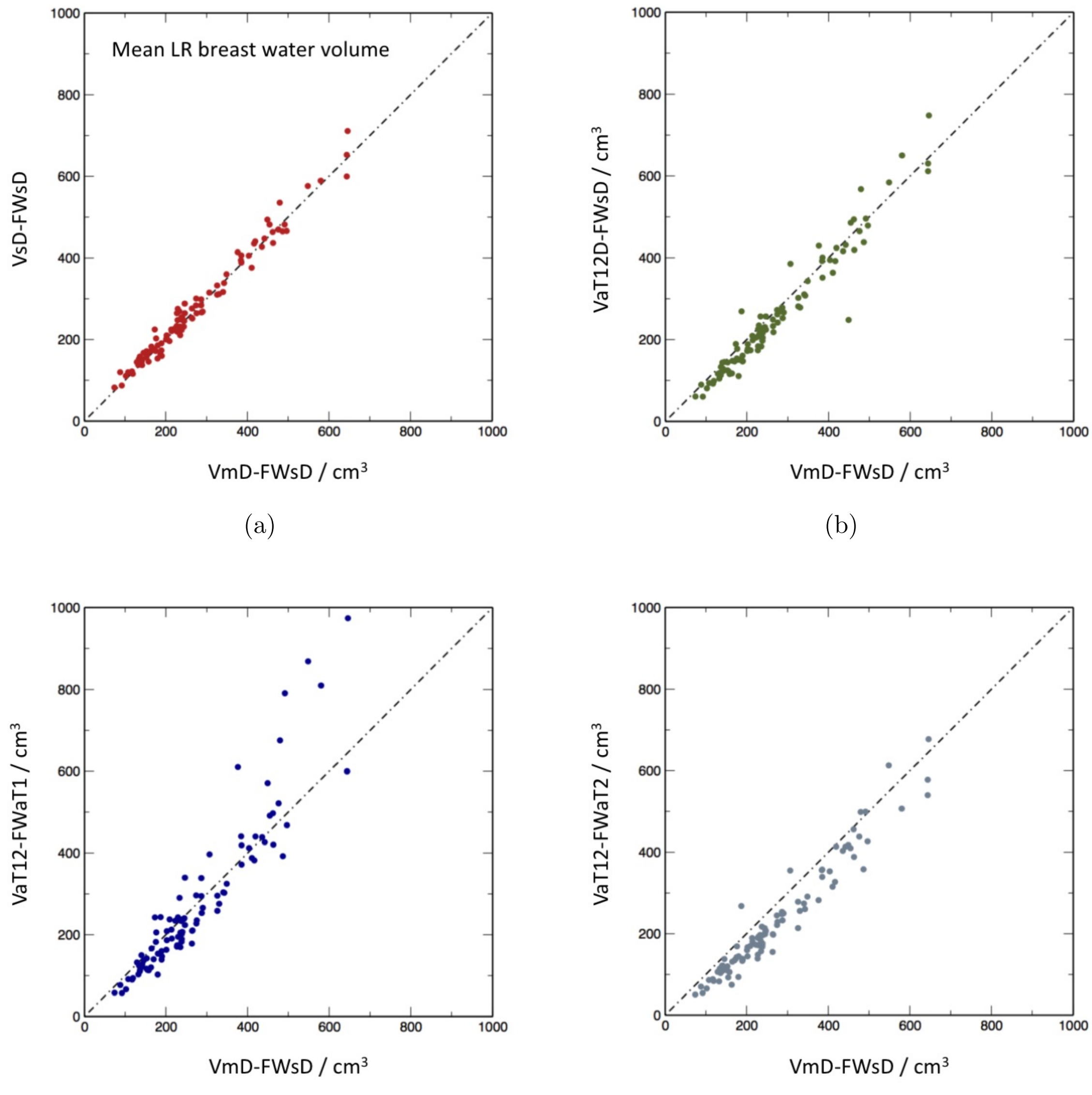


FIG. 9: Scatter plots of mean left and right breast water volumes in cm^3 for the different methods in comparison to VmD-FWsD. For nomenclature see caption to Figure 8.

1 cm increase in height, the analysis methods gave the following relative increases in breast volume: VmD 1.05[0.98, 1.11], VsD 1.04[0.98,1.11], VaT12D-FWsD was 1.05[0.97, 1.12] , VaT12-FWaT1 1.05[095, 1.03], VaT12-FWT2 1.05[0.95, 1.13]. However, height was not associated with breast water fraction.

No associations were found with any of: age of menarche, use of oral contraception, smoking, alcohol intake or maternal mammographic density.

From the similarity of all these statistics, we conclude that the exact details of the seg-

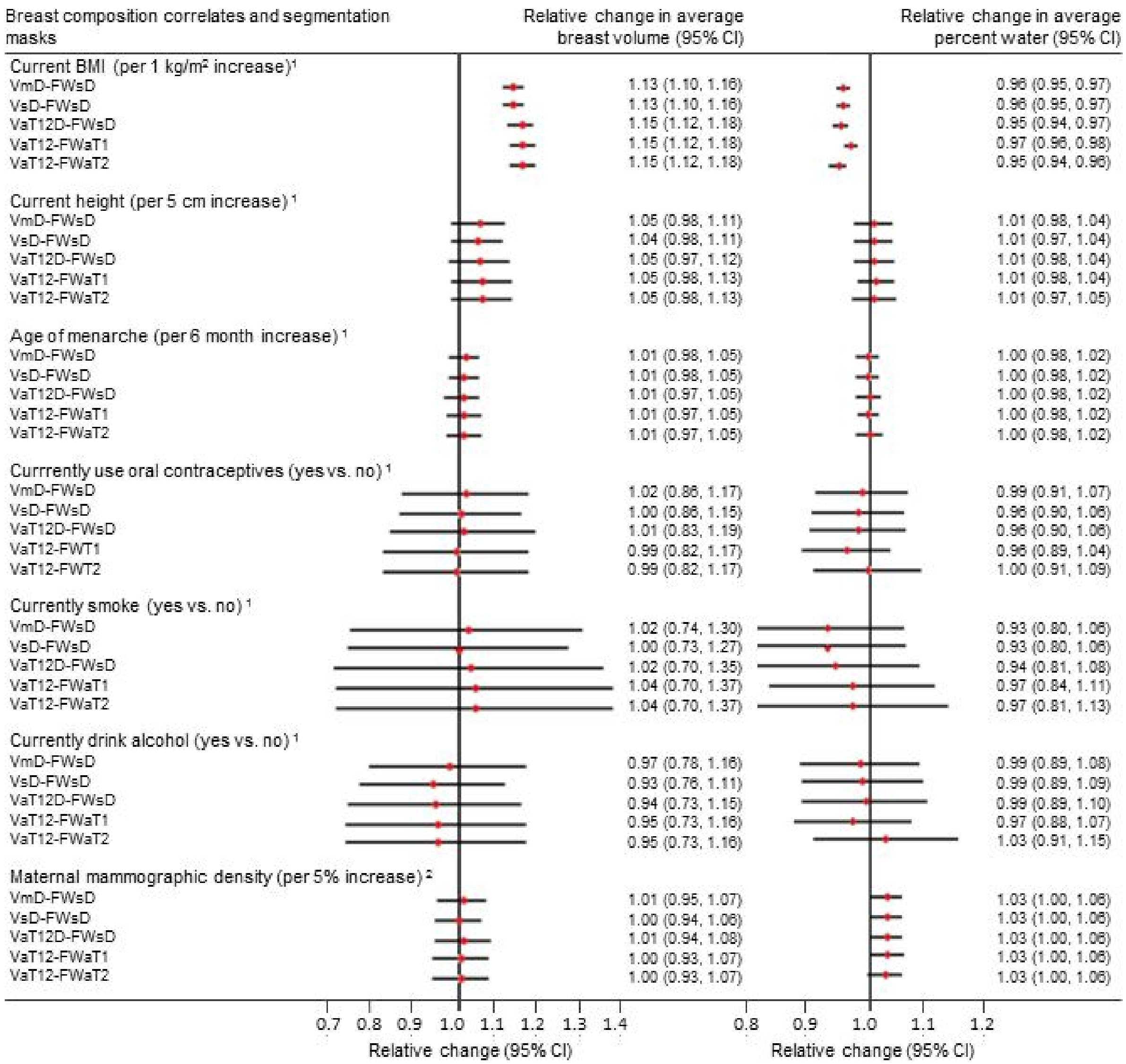


FIG. 10: Results of epidemiological analysis. For further details, see the supplementary information.

mentation methods are not significant at the level of this cohort analysis.

IV. DISCUSSION

Our results show that, as in many segmentation problems, the degree of success of the automated algorithms varies significantly between subjects. Figure 5 and Table II demonstrate excellent performance by all of the algorithms, whereas the degree of correspondence with the expert manual segmentation is considerably poorer in Figure 6 and Table III. However, it should be noted that even the expert human observer is less able to provide a good repeat segmentation.

The ICCs for total breast volume in Table IV demonstrate good agreement between all methods, but interestingly, slightly closer agreement between VaT12 and the two Dixon-based methods (VmD or VsD) than between VaT12D and the Dixon methods. As described above, VaT12D is created by simply resampling VaT12 in the Dixon coordinate space, which has a coarser slice thickness, using appropriate blurring and nearest neighbour interpolation. Although movement between the Dixon and T1w or T2w scans could explain this disparity, registering the volumes did not improve the results. The resampling process appears to amplify the difference between VaT12 and VmD or VsD, but we have not analysed this further, given that it is a relatively small effect.

It would, of course, be interesting to compare the output of the VaT1T2 method directly with manual segmentation of the high-resolution T1w dataset in its native reference frame, without the need to down-sample. However, the workload involved in creating high-resolution manual segmentations is prohibitive. In the Supplementary Information, we report anecdotal results for five such cases with full high-resolution manual segmentations.

Also of note from comparison of the scatter-plots of Figure 7 is that each of methods VsD, VaT12D and VaT12 increasingly over-estimates the breast volume in comparison to VmD as the mean left and right breast size increases. This is most apparent for VaT12. The trend to larger error is, of course logical – similar percentage errors between the methods will result in greater absolute differences the larger the breast – but it is not currently clear why all methods are biased to *over*-estimate the volume in this region. Method VaT12D also *under*-estimates the breast volume for smaller breasts compared with the manual segmentation VmD and the reason for this, too, is unclear.

The biggest discrepancy between analysis methods, as shown by the scatter plots, is in the assessment of mean breast water volume (and, hence, water fraction — data not shown). The VsD-FWsD and VaT12-FWsD methods both use Dixon source data and differ from VmD-FWsD only via the breast outline previously described. The methods all give very similar results (ICCs 0.995 and 0.992 in Table V). By contrast, the correlation between the Dixon-based VmD-FWsD and VaT12-FWaT1 is weaker, and the VaT12-FWaT2 result additionally shows a bias (Figure 8). However, it is important to note that the assumption that water fractions based on the Dixon method can be regarded as a gold standard for true parenchymal fraction is much less compelling than the previous assumption that VmD is the gold-standard volume. We justify our choice of VmD-FWsD as the method of comparison on the basis that it is consistent with previous work in the field[49] (and indeed an improvement), but Ledger *et al.*[52] have demonstrated that there is a significant degree of variability between different Dixon-based methods, depending on the exact design of the pulse sequence. It is unsurprising that a segmentation based on a completely different MRI contrast mechanism should be less highly correlated. What is nevertheless highly encouraging is that the correlation remains as strong as it is — the worst value reported in Table V is 0.920 — and this suggests that the use of MRI as a modality will prove to be a robust choice for breast analysis.

[A salutory lesson from the scatter graphs is the constant need for vigilance and appropriate quality control when processing large cohorts of data. During the review of this paper a referee noticed an outlier, which turned out to be the result of an easily-corrected error that caused the mask for the entire right breast to be missing. Such "edge" cases, occurring very infrequently, remain a significant challenge in the adoption of automated pipelines. Any requirement for manual inspection of each dataset to check the output negates to some extent the advantages of fully-automated segmentation processes, and an appropriate balance needs to be determined for each application.]

Another feature highlighted by all of these results is the problem inherent in the use of quantitative metrics such as Dice and correlation coefficients, which (despite their apparent calculation "accuracy") are a very blunt tool for analysing a complex situation. Are all of the voxels that fail to overlap equally important? Is much of the difference between the observer and the automated methods in fact caused by the choice of how much of the axilla is included and is this region of any significance biologically?

A first reading of the coefficients presented here suggests that the VsD breast outline segmentation, followed by the FWsD tissue segmentation method is the best-performing of the computer-aided tools presented here. But is it the most suitable? Ultimately, the choice of segmentation method needs to weigh up the following points:

- To what extent does the application demand a segmentation that is as good as that of an expert radiologist? Two extremes here might be the planning of radiotherapy treatment for an individual patient, where high correspondence is vital, and the calculation of epidemiological parameters for a Big Data cohort, where errors might well "average out."

- To what extent is the ground truth knowable? For a given set of intra- and inter-observer performance metrics evaluated on a test cohort, what performance thresholds should be regarded as "acceptable" for automated segmentations?

- How widely available are the required source data? As previously noted, the Dixon protocol is not routinely included in clinical examinations, thus limiting the applicability of breast density measurements based on the VsD-FWsD method.

- How robust is the method?

- To what extent are speed, convenience and consistency of method to be preferred over accuracy?

In our case, consideration of all of the above led to the use of the VaT12 method, rather than VsD, for segmentation of the remaining 300 cases in the cohort (results not presented). This choice was made largely on the basis of improved automation and on the epidemiological evidence from the 200-strong training and test datasets, as described in Section III C, where key epidemiological parameters were found to be identical, within confidence limits, for both methods.

V. CONCLUSION

We have presented what we believe to be the first detailed comparison on a large, population-based cohort of two methods of breast-outline segmentation based on completely different approaches. These have been coupled with two methods of fat-water discrimination based on fundamentally different MR contrast mechanisms. All combinations of the methods studied are in very strong agreement, as seen both visually and via inter-class correlation coefficients, and are suitable for large-scale epidemiological analysis. We have discussed the assumptions behind the methods and posed a number of general questions that we believe need to be answered each time a decision is made on whether and how to perform automated segmentation.

ACKNOWLEDGMENTS

We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. In particular we would like to thank study nurses, Elizabeth Folkes and Sally Pearce, and CRIC radiographer, Aileen Wilson, for performing MRI acquisitions of all the participants. The UK Medical Research Council and the Wellcome Trust (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC.

Authors SJD and MOL acknowledge CRUK and EPSRC support to the Cancer Imaging Centre at ICR and RMH in association with MRC and Department of Health C1060/A10334, C1060/A16464 and NHS funding to the NIHR Biomedical Research Centre and the Clinical Research Facility in Imaging. Authors JH, BE and DH were funded by the European 7th Framework Program grants VPH-PRISM (FP7-ICT-2011-9, 601040), VPH-PICTURE (FP7-ICT-2011-9, 600948) and the Engineering and Physical Sciences Research Council grant MIMIC (EP/K020439/1). IdSS was supported by funding from Cancer Research UK (grant number C405/A12730).

DISCLOSURE

The authors are not aware of any conflicts of interest.

REFERENCES

[1]V. A. McCormack and I. D. S. Silva, "Breast density and parenchymal patterns as markers of breast cancer risk: A meta-analysis," Cancer Epidemiology Biomarkers & Prevention **15**, 1159–1169 (2006).

[2]E. Vilaprinyo, C. Forne, M. Carles, M. Sala, R. Pla, X. Castells, L. Domingo, M. Rue, and I. S. G. Interval Canc, "Cost-effectiveness and harm-benefit analyses of risk-based screening strategies for breast cancer," Plos One **9** (2014), 10.1371/journal.pone.0086858.

[3]E. R. Price, A. W. Keedy, R. Gidwaney, E. A. Sickles, and B. N. Joe, "The potential impact of risk-based screening mammography in women 40-49 years old," American Journal of Roentgenology **205**, 1360–1364 (2015).

[4]S. Ciatto, N. Houssami, A. Apruzzese, E. Bassetti, B. Brancato, F. Carozzi, S. Catarzi, M. P. Lamberini, G. Marcelli, R. Pellizzoni, B. Pesce, G. Risso, F. Russo, and A. Scorsolini, "Categorizing breast mammographic density: intra- and interobserver reproducibility of bi-rads density categories," Breast **14**, 269–275 (2005).

[5]R. Highnam, S. M. Brady, M. J. Yaffe, N. Karssemeijer, and J. Harvey, "Robust breast composition measurement - volparatm," in *Digital Mammography: 10th International Workshop, IWDM 2010, Girona, Catalonia, Spain, June 16-18, 2010. Proceedings*, edited by J. Mart, A. Oliver, J. Freixenet, and R. Mart (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010) pp. 342–349.

[6]G. Waade, R. Highnam, I. Hauge, M. McEntee, S. Hofvind, E. Denton, J. Kelly, J. Sarwar, and P. Hogg, "Error in recorded compressed breast thickness measurement impacts on volumetric density classification using volpara v1.5.0 software," Medical Physics **In press** (2016).

[7]A. Gubern-Mérida, M. Kallenberg, B. Platel, R. M. Mann, R. Mart, and N. Karssemeijer, "Volumetric breast density estimation from full-field digital mammograms: A validation study," PLoS ONE **9**, e85952 (2014).

[8]D. J. Thompson, M. O. Leach, G. Kwan-Lim, S. A. Gayther, S. J. Ramus, I. Warsi, F. Lennard, M. Khazen, E. Bryant, S. Reed, *et al.*, "Assessing the usefulness of a novel mri-based breast density estimation algorithm in a cohort of women at high genetic risk of breast cancer: the uk maribs study," Breast Cancer Research **11**, R80 (2009).

[9]P. Hayton, P. Hayton, J. M. Brady, J. M. Brady, L. Tarassenko, L. Tarassenko, N. Moore,

and N. Moore, "Analysis of dynamic mr breast images using a model of contrast enhancement," Medical Image Analysis **1**, 207–24 (1997).

[10]T. Twellmann, O. Lichte, and T. W. Nattkemper, "An adaptive tissue characterization network for model-free visualization of dynamic contrast-enhanced magnetic resonance image data," Medical Imaging, IEEE Transactions on **24**, 1256–1266 (2005).

[11]M. Koenig, "Automatic cropping of breast regions for registration in mr mammography," Proceedings of SPIE **5747**, 1563–1570 (2005).

[12]J. Yao, "Classification and calculation of breast fibroglandular tissue volume on spgr fat suppressed mri," Proceedings of SPIE **5747**, 1942–1949 (2005).

[13]W. L. W. Lu, J. Y. J. Yao, C. L. C. Lu, S. Prindiville, and C. Chow, "Dce-mri segmentation and motion correction based on active contour model and forward mapping," Seventh ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD'06) , 0–4 (2006).

[14]V. Giannini, A. Vignati, L. Morra, D. Persano, D. Brizzi, L. Carbonaro, A. Bert, F. Sardanelli, and D. Regge, "A fully automatic algorithm for segmentation of the breasts in dce-mr images," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE* (IEEE) pp. 3146–3149.

[15]L. Wang, B. Platel, T. Ivanovskaya, M. Harz, H. K. Hahn, and Ieee, "Fully automatic breast segmentation in 3d breast mri," 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI) , 1024–1027 (2012).

[16]S. Wu, S. Weinstein, and D. Kontos, "Atlas-based probabilistic fibroglandular tissue segmentation in breast mri," Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention **15**, 437–45 (2012).

[17]S. Wu, S. P. Weinstein, E. F. Conant, A. R. Localio, M. D. Schnall, and D. Kontos, "Fully automated chest wall line segmentation in breast mri by using context information," in *Medical Imaging 2012: Computer-Aided Diagnosis*, Proceedings of SPIE, Vol. 8315, edited by B. VanGinneken and C. L. Novak (2012).

[18]S. Wu, S. P. Weinstein, E. F. Conant, and D. Kontos, "Automated fibroglandular tissue segmentation and volumetric density estimation in breast mri using an atlas-aided fuzzy c-means method," Medical Physics **40** (2013).

[19]S. Wu, S. P. Weinstein, E. F. Conant, and D. Kontos, "Fully-automated fibroglandular

tissue segmentation and volumetric density estimation in breast mri by integrating a continuous max-flow model and a likelihood atlas," in *Medical Imaging 2013: Computer-Aided Diagnosis*, Proceedings of SPIE, Vol. 8670, edited by C. L. Novak and S. Aylward (2013).

[20] A. Gubern-Merida, M. Kallenberg, R. Marti, and N. Karssemeijer, "Multi-class probabilistic atlas-based segmentation method in breast mri," in *Pattern Recognition and Image Analysis: 5th Iberian Conference, Ibpria 2011*, Lecture Notes in Computer Science, Vol. 6669, edited by J. Vitria, J. M. Sanches, and M. Hernandez (2011) pp. 660–667.

[21] A. Gubern-Mérida, M. Kallenberg, R. Mart, and N. Karssemeijer, "Segmentation of the pectoral muscle in breast mri using atlas-based approaches," in *Medical Image Computing and Computer-Assisted InterventionMICCAI 2012* (Springer, 2012) pp. 371–378.

[22] A. Gubern-Mérida, M. Kallenberg, R. Mann, R. Marti, and N. Karssemeijer, "Breast segmentation and density estimation in breast mri: A fully automatic framework," IEEE Journal of Biomedical and Health Informatics **19**, 349–357 (2015).

[23] C. Gallego-Ortiz and A. Martel, "Automatic atlas-based segmentation of the breast in mri for 3d breast volume computation," Medical physics **39**, 5835–5848 (2012).

[24] F. Khalvati, C. Gallego-Ortiz, S. Balasingham, and A. L. Martel, "Automated segmentation of breast in 3-d mr images using a robust atlas," IEEE Transactions on Medical Imaging **34**, 116–125 (2015), 0.

[25] C. Gallego and A. L. Martel, "Automatic model-based 3d segmentation of the breast in mri," in *Conference on Medical Imaging 2011 - Image Processing*, Proceedings of SPIE, Vol. 7962 (2011).

[26] G. Ertas, H. O. Gulcur, M. Tunaci, and M. Dursun, "k-means based segmentation of breast region on mr mammograms," Magnetic Resonance Materials in Physics, Biology and Medicine **19**, 317 (2006).

[27] G. Ertas, H. O. Gulcur, M. Tunaci, O. Osman, and O. N. Ucan, "A preliminary study on computerized lesion localization in mr mammography using 3d nmitr maps, multilayer cellular neural networks, and fuzzy c-partitioning," Medical physics **35**, 195–205 (2008).

[28] C.-M. Wang, X.-X. Mai, G.-C. Lin, and C.-T. Kuo, "Classification for breast mri using support vector machine," 8th IEEE International Conference on Computer and Information Technology Workshops: Cit Workshops 2008, Proceedings , 362–367 (2008).

[29] Y. Wang, G. Morrell, M. E. Heibrun, A. Payne, and D. L. Parker, "3d multi-parametric breast mri segmentation using hierarchical support vector machine with coil sensitivity

correction," Academic Radiology **20**, 137–147 (2013).

[30] C. Klifa, J. Carballido-Gamio, L. Wilmes, A. Laprie, C. Lobo, E. DeMicco, M. Watkins, J. Shepherd, J. Gibbs, and N. Hylton, "Quantification of breast tissue index from mr data using fuzzy clustering," in *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, Vol. 1 (IEEE) pp. 1667–1670.

[31] C. Klifa, J. Carballido-Gamio, L. Wilmes, A. Laprie, J. Shepherd, J. Gibbs, B. Fan, S. Noworolski, and N. Hylton, "Magnetic resonance imaging for secondary assessment of breast density in a high-risk cohort," Magnetic Resonance Imaging **28**, 8–15 (2010).

[32] S.-C. Yang, C.-M. Wang, H.-H. Hsu, P.-C. Chung, G.-C. Hsu, C.-J. Juan, and C.-S. Lo, "Contrast enhancement and tissues classification of breast mri using kalman filter-based linear mixing method," Computerized Medical Imaging and Graphics **33**, 187–196 (2009).

[33] K. Nie, J.-H. Chen, S. Chan, M.-K. I. Chau, H. J. Yu, S. Bahri, T. Tseng, O. Nalcioglu, and M.-Y. Su, "Development of a quantitative method for analysis of breast density based on three-dimensional breast mri," Medical Physics **35** (2008).

[34] A. Sathya, S. Senthil, and A. Samuel, "Segmentation of breast mri using effective fuzzy c-means method based on support vector machine," Proceedings of the 2012 World Congress on Information and Communication Technologies , 67–72 (2012).

[35] M. Lin, S. Chan, J.-H. Chen, D. Chang, K. Nie, S.-T. Chen, C.-J. Lin, T.-C. Shih, O. Nalcioglu, and M.-Y. Su, "A new bias field correction method combining n3 and fcm for improved segmentation of breast density on mri," Medical Physics **38**, 5–14 (2011).

[36] M. Lin, J.-H. Chen, X. Wang, S. Chan, S. Chen, and M.-Y. Su, "Template-based automatic breast segmentation on mri by excluding the chest region," Medical Physics **40** (2013).

[37] G. Ertas, S. J. Doran, and M. O. Leach, "A computerized volumetric segmentation method applicable to multicentre mri data to support computer aided breast tissue analysis, density assessment and lesion localization," Medical and Biological Engineering and Computing , 1–12 (2016).

[38] W. T. Dixon, "Simple proton spectroscopic imaging," Radiology **153**, 189–194 (1984).

[39] P. H. England, "Breast screening: professional guidance," (2016).

[40] A. Boyd, J. Golding, J. Macleod, D. A. Lawlor, A. Fraser, J. Henderson, L. Molloy, A. Ness, S. Ring, and G. Davey Smith, "Cohort profile: the 'children of the 90s' – the index offspring of the avon longitudinal study of parents and children." Int J Epidemiol **42**, 111–127 (2013).

[41] http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary.

[42] D. L. Pham and J. L. Prince, "An adaptive fuzzy c-means algorithm for image segmentation in the presence of intensity inhomogeneities," Pattern Recognition Letters **20**, 57–68 (1999).

[43] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: Improved n3 bias correction," Ieee Transactions on Medical Imaging **29**, 1310–1320 (2010).

[44] N. L. of Medicine, "Insight segmentation and registration toolkit (itk)," http://www.itk.org/.

[45] A. Gubern-Merida, M. Kallenberg, R. Marti, and N. Karssemeijer, "Segmentation of the pectoral muscle in breast mri using atlas-based approaches," Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention **15**, 371–8 (2012).

[46] L. Wang, K. Filippatos, O. Friman, and H. K. Hahn, "Fully Automated Segmentation of the Pectoralis Muscle Boundary in Breast MR Images," in *MEDICAL IMAGING 2011: COMPUTER-AIDED DIAGNOSIS*, Proceedings of SPIE, Vol. 7963, edited by Summers, RM and VanGinneken, B (2011).

[47] A. Gubern-Mérida, L. Wang, M. Kallenberg, R. Marti, H. K. Hahn, and N. Karssemeijer, "Breast segmentation in mri: quantitative evaluation of three methods," Medical Imaging 2013: Image Processing **8669** (2013).

[48] L. D. Griffin, "The second order local-image-structure solid," IEEE Transactions on Pattern Analysis and Machine Intelligence **29**, 1355–1366 (2007).

[49] C. S. Poon, M. J. Bronskill, R. M. Henkelman, and N. F. Boyd, "Quantitative magnetic resonance imaging parameters and their relationship to mammographic pattern," Journal of the National Cancer Institute **84**, 777–781 (1992).

[50] H. Song, X. Cui, and F. Sun, "Breast tissue 3d segmentation and visualization on mri," International journal of biomedical imaging **2013**, 859746–859746 (2013).

[51] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classification of mr images of the brain," IEEE Transactions on Medical Imaging **18**, 897–908 (1999).

[52] A. E. W. Ledger, E. D. Scurr, J. Hughes, A. Macdonald, T. Wallace, K. Thomas, R. Wilson, M. O. Leach, and M. A. Schmidt, "Comparison of dixon sequences for estimation of percent

breast fibroglandular tissue," PLoS ONE **11**, e0152152 (2016).