Published in final edited form as:
Med Image Anal. 2014 January ; 18(1): . doi:10.1016/j.media.2013.09.001.

A Collaborative Resource to Build Consensus for Automated Left Ventricular Segmentation of Cardiac MR Images

Avan Suinesiaputra[a], **Brett R. Cowan**[a], **Ahmed O. Al-Agamy**[d], **Mustafa A. AlAttar**[f], **Nicholas Ayache**[c], **Ahmed S. Fahmy**[d,e], **Ayman M. Khalifa**[f,j], **Pau Medrano-Gracia**[b], **Marie-Pierre Jolly**[i], **Alan H. Kadish**[g], **Daniel C. Lee**[g], **Ján Margeta**[c], **Simon K. Warfield**[h], and **Alistair A. Young**[a]

[a]Department of Anatomy with Radiology, University of Auckland, New Zealand [b]Auckland Bioengineering Institute, University of Auckland, New Zealand [c]Asclepios Research Project, INRIA Sophia-Antipolis, France [d]Center for Informatics Science, Nile University, Cairo, Egypt [e]Systems and Biomedical Engineering, Cairo University, Cairo, Egypt [f]Diagnosoft Inc., Cardiac Image Analysis, Morrisville, NC, USA [g]Division of Cardiology, Northwestern University, USA [h]Computational Radiology Laboratory, Harvard Medical School, USA [i]Siemens Corporation, Corporate Technology, Imaging and Computer Vision, Princenton, NJ, USA [j]Biomedical Engineering Department, Helwan University, Cairo, Egypt

Abstract

A collaborative framework was initiated to establish a community resource of ground truth segmentations from cardiac MRI. Multi-site, multi-vendor cardiac MRI datasets comprising 95 patients (73 men, 22 women; mean age 62.73 ± 11.24 years) with coronary artery disease and prior myocardial infarction, were randomly selected from data made available by the Cardiac Atlas Project (Fonseca et al., 2011). Three semi- and two fully-automated raters segmented the left ventricular myocardium from short-axis cardiac MR images as part of a challenge introduced at the STACOM 2011 MICCAI workshop (Suinesiaputra et al., 2012). Consensus myocardium images were generated based on the Expectation-Maximization principle implemented by the STAPLE algorithm (Warfield et al., 2004). The mean sensitivity, specificity, positive predictive and negative predictive values ranged between 0.63-0.85, 0.60-0.98, 0.56-0.94 and 0.83-0.92, respectively, against the STAPLE consensus. Spatial and temporal agreement varied in different amounts for each rater. STAPLE produced high quality consensus images if the region of interest was limited to the area of discrepancy between raters. To maintain the quality of the consensus, an objective measure based on the candidate automated rater performance distribution is proposed. The consensus segmentation based on a combination of manual and automated raters were more consistent than any particular rater, even those with manual input. The consensus is expected to improve with the addition of new automated contributions. This resource is open for future contributions, and is available as a test bed for the evaluation of new segmentation algorithms, through the Cardiac Atlas Project (www.cardiacatlas.org).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Cardiac magnetic resonance (CMR) imaging has been widely adopted in many clinical institutions to routinely diagnose patients with cardiac diseases (Ishida et al., 2009; Muzzarelli et al., 2011; Karamitsos and Myerson, 2011). CMR is non-ionising, non-invasive and provides a range of contrast mechanisms. This makes it a versatile diagnostic technique that enables multiple protocols, e.g. cine functional studies, tissue characterization, perfusion, stress imaging, velocity and flow, within a single session (Pennell, 2010). In particular, CMR is widely acknowledged as the most accurate method for the left ventricular (LV) mass and volumes calculation. Due to its non-invasive and well validated accuracy, CMR is now being used in several large epidemiological studies including MESA (Bild et al., 2002) and the UK Biobank (Ollier et al., 2005). There is therefore a pressing need for robust fully automated segmentation of LV myocardium in the CMR domain.

However, automated delineation of myocardial boundaries remains a difficult problem, due to artifacts arising from flow, motion, off-resonance behavior and noise. Smooth intensity gradients around the myocardium create different opinions even among expert analysts (Paetsch et al., 2006). Papillary muscles, trabeculae and intensity inhomogeneities also contribute to this problem. These problems give rise to differences in ground truth delineations arising from different observers. Figure 1 shows examples of the disagreements among expert analysts, particularly on papillary muscle (Fig. 1(a)), trabeculae structures (Fig. 1(b)), and LV outflow tract (Fig. 1(c)).

The difficulty to delineate myocardial borders has been a major challenge for researchers to develop automated segmentation algorithms for more than a decade. Different approaches of automatic CMR segmentation methods have been proposed (see surveys in Petitjean and Dacher, 2011; Frangi et al., 2001). Qualitative and quantitative segmentation results have been presented. However, few of these methods have been compared on substantial datasets using an objective comparison technique. Each published method conducted experiments with private image data-bases and different reference contour definitions. This makes objective comparisons between methods difficult to perform.

In a recent survey of automated segmentation methods from cardiac MRI (Petitjean and Dacher, 2011), the lack of publicly available image datasets and a common performance evaluation protocol have been outlined as an open problem in LV segmentation. In the absence of an exact ground truth, a consensus method must be employed to provide reference segmentations for the evaluation of automated methods (Warfield et al., 2004). A consensus ground truth for myocardium segmentation should therefore be established in the CMR research community using widely available images in order to enable proper benchmarking and validation. Similar resources have been made available in other domains, for example in the segmentation of the carotid arteries for stenosis evaluation (Schaap et al., 2009; Hameeteman et al., 2011), the detection of pulmonary nodules from lung CT images (van Ginneken et al., 2010; Murphy et al., 2011), airway tree segmentation (Lo et al., 2012), and brain image segmentation (Shattuck et al., 2009).

In the cardiac domain, such a benchmarking resource has only been available for small datasets at a few cardiac frames. For instance, the 2009 MICCAI LV segmentation challenge included 45 cases at end-diastole and end-systole (Radau et al., 2009). The main problem is that collecting many manually drawn expert contours from a large number of cases, defined on all slices and cardiac frames, is physically impractical. Expert segmentation is a time consuming and painstaking process which involves a considerable resource expense. To date, there have therefore been no large datasets with validated ground truth available for this purpose. To solve this problem, we initiated a collaborative framework to establish common consensus myocardium images that were estimated by combining the knowledge of human observer expertise with the objectivity of automated segmentation methods. The approach we have adopted is to leverage a variety of published and validated semiautomatic and automatic methods applied to a widely available, clinically important dataset in order to develop a consensus ground truth. Hence, our contributions in this paper are the following:

1. To establish a community resource of ground truth images based on common data for the development, validation and benchmarking of LV segmentation algorithms.
2. To present a pipeline for building consensus myocardium images by involving independent multi-center automated methods, which allows the inclusion of automated results into the consensus under certain conditions, so that the consensus will be iteratively refined as new centers participate.
3. To demonstrate that the consensus images built using this framework are more consistent than any particular rater, even those with manual input.
4. To quantify for the first time the regional and temporal variation in agreement between raters in the left ventricle segmentation application domain.

In this initial study, five raters were involved: three which included manual input (one which defined contours through an interactive customization of a 3D+t LV shape representation by an experienced operator and two which were initialised by manually drawn contours) and two fully-automated raters. These represent all groups who participated in the 2011 LV Segmentation Challenge held at the MICCAI 2011 workshop Statistical Atlases and Computational Models of the Heart (STACOM) Suinesiaputra et al. (2012).

To estimate the consensus images, we adapted an Expectation Maximization (EM) based algorithm (STAPLE) for cardiac MRI. STAPLE collates evidence from each rater to estimate the best possible segmentations by maximizing the performance of all raters (Warfield et al., 2004). A useful feature of the STAPLE algorithm is its well-founded formulation to estimate true segmentation images from the input raters using objective evaluation criteria. This is inherited from the basic principles of EM algorithm (Dempster et al., 1977). STAPLE has been applied successfully in the validation of human brain images (Liu et al., 2007; Archip et al., 2007), consensus guidelines in prostate MRI (Hwee et al., 2011) and observer reliability study in pelvic MRI (Hoyte et al., 2011). In cardiac applications, a preliminary work to identify cardiac landmarks by using STAPLE was reported in (Xing et al., 2011).

This paper describes the framework for the consensus ground truth estimation process, the resources established for the benchmarking and validation of automated segmentation techniques, mechanisms for the inclusion of automated results into the consensus, and the results obtained so far. To maintain the consensus images quality, a new set of criteria based on clinical LV function and performance distribution is proposed for the inclusion of automated raters. The resource is open for contributions on an ongoing basis, since we expect that the consensus will become more robust with the addition of more raters. The consensus segmentations are available on request from the Cardiac Atlas Project website[1].

In the following section, we describe the infrastructure that we established for this study. Brief explanations of each rater that contributed to this initial consensus estimation are presented in Section 3. A short description about the STAPLE algorithm is presented in Section 4 as the background for the consensus estimation framework. In Section 5, we describe the appropriate prior estimation approach for the STAPLE algorithm in the context

[1] http://www.cardiacatlas.org

of cardiac MRI. Section 6 defines the performance metrics used to evaluate the consensus results, as well as an objective measure for the inclusion of a new rater to the consensus. Section 7 describes validation experiments comparing different variations of consensus. Section 8 presents the results. Section 9 discusses the characteristics of each rater and the effects of adding fully automated raters to the consensus images, followed by conclusion in Section 10.

2. Data

Two hundred patients were randomly selected from the DETERMINE cohort (Defibrillators To Reduce Risk by Magnetic Resonance Imaging Evaluation) (Kadish et al., 2009). One hundred were made available as training data, with manual segmentation, and the other hundred were reserved for validation. Of these, five studies could not be processed by all raters due to inconsistencies in the image parameters, e.g. a different image field of view in one series, leaving 95 cases available for consensus. The DETERMINE study comprises of patients with coronary artery disease and regional wall motion abnormalities due to prior myocardial infarction. This is a clinically important patient group since mass and volume are important diagnostic and prognostic indicators of adverse remodeling. Studies were acquired at multiple sites using multiple scanner vendors. The data were made available through the Cardiac Atlas Project (Fonseca et al., 2011). Characteristics of the patient data are shown in Table 1.

The CMR images were based on the steady-state free precession (SSFP) pulse sequence. CMR parameters varied between cases giving a heterogenous mix of scanner types and imaging parameters. MR scanner systems were GE Medical Systems (Signa 1.5T), Philips Medical Systems (Achieva 1.5T, 3.0T, and Intera 1.5T), and Siemens (Avanto 1.5T, Espree 1.5T and Symphony 1.5T). Typical short-axis slice parameters were either a 6 mm slice thickness with 4 mm gap or 8 mm slice thickness with 2 mm gap. Image size was ranging from 138 × 192 to 512 × 512 pixels. The temporal resolution was between 19 and 30 frames. Long axis images in the four and two chamber orientations were also available.

Basal, mid-ventricle and apical slices were defined semiautomatically by using one of the manual raters, which employed a finite element LV model. The LV model was divided for each patient into three regions with equal height and slices were assigned into regions on the basis of this division. Note that the consensus did not make any distinction between slices; this division was only performed for the evaluation of the results.

3. Contributing raters

There were two fully-automated raters (SCR and INR) and three semi-automated raters with manual input (AO, AU and DS) who participated to this study. A brief summary of each rater segmentation method is given in the following subsections.

3.1. Deformable registration method (SCR)

This fully automatic algorithm segments all phases in one slice at a time using deformable registration, taking advantage of the strong temporal correlation between phases. The main idea of this algorithm is to use an inverse consistent deformable registration to register all frames to the first frame in one slice. Then, the segmentation can be applied to any frame and propagated to any other frame in the sequence through the forward and backward deformation fields.

First, the LV blood pool is automatically detected based on the moving components using the first harmonic of the Fourier transform over time in each slice (Jolly, 2008).

Subsequently, connected components between slices are grouped using isoperimetric clustering to form the 3D blood pool. When long axis slices are available, they are used to generate a plane approximation for the mitral valve by using a machine learning algorithm, which detects the mitral valve leaflet anchor points in ED and ES frames (Lu et al., 2010).

The LV segmentation method is based on an inverse consistent deformable registration approach (Jolly et al., 2010). The registration method (Guetter et al., 2011) computes a dense deformation field between any two frames in a slice without having to explicitly register every possible pair of frames. This is achieved by making the registration inverse consistent so that forward and backward deformation fields are recovered during the registration of all frames to an arbitrary keyframe, e.g. frame 1. The deformation field between frames *i* and *j* is obtained by composing the deformation field between frames 1 and *j* and the inverse deformation field between frames 1 and *i*.

The core of the algorithm is illustrated in Fig. 2. For each slice, the first frame is segmented by recovering endocardium and epicardium contours using a shortest path algorithm in polar space. The contours are then propagated to the other frames using the deformation fields and cost function is evaluated. This process is repeated for all frames in the slice and the segmentation that results in the smallest cost function is retained. These contours are transferred to the next slice and used as priors to the contour-based segmentation. More details of this approach are presented in (Jolly et al., 2012). No user input was required for this method.

3.2. Layered spatio-temporal forests algorithm (INR)

An automatic machine learning based method was proposed in (Margeta et al., 2012) to tackle the segmentation problem by using only the class labels provided without prior knowledge, such as a statistical model or circularity of the ventricle. This method extends the previous random forest segmentation algorithms for multiple sclerosis lesions from multi-channel MRIs (Geremia et al., 2011) and for LV from 3D ultrasound images (Lempitsky et al., 2009). The images were treated directly as 3D+t volumes and the segmentation problem was defined as voxel-wise classification into myocardium and background.

For each voxel, a large number of spatio-temporal regional average intensity differences were generated by randomly varying size and position of the regions with respect to the tested voxel (see Fig. 3). This resulted in a large number of possible measurements from which only the ones relevant for segmentation of the left ventricle were selected. This relevancy was determined by the information gain criteria and was used to determine optimal parameters of the splitting function at each node (e.g. feature type, box positions, size and threshold). By recursively splitting the data and selecting the optimal split parameters from a random sample of the parameter subset, a tree-based representation of the segmentation problem was constructed during the training. Class label probability distributions were assigned to each leaf based on the class distributions of the voxels that are falling into that leaf.

During the classification process, the voxel is passed through the nodes in all trees based on the pre-trained split parameters and its final posterior probability class distribution is obtained as the average distribution of all reached leaf nodes (see Criminisi et al., 2012, for more details on random forests).

The algorithm used two classification layers to tackle the segmentation problem. Both layers were trained to segment the LV myocardium, each for a different purpose. The probability map from the first layer was used to correct the cardiac sequences for acquisition pose differences by applying a robust block matching algorithm (Ourselin et al., 2000) directly on the probability maps. Furthermore, the probability map was used to estimate myocardial intensity for MRI intensity standardisation (Nyúl and Udupa, 1999). The second layer was then retrained on intensity and pose standardized images, added absolute voxel coordinates as features and was used for a more accurate final segmentation.

To be used in this collation study, the second layer posterior probability maps were reoriented to the original pose and each voxel was then assigned the class label (myocardium or background) with the highest posterior probability, resulting binary images of the myocardium. This method required no user input.

3.3. Contour-constrained optical flow tracking (AO)

In this approach, a modified optical flow (OF) algorithm was used to track an initial contour manually drawn on the initial timeframe. Let $P_s^t = \left(x_s^t, y_s^t\right)$ be the *s*th contour point at timeframe *t*, that moves with a displacement of $\left(u_s^t, v_s^t\right)$ to the next timeframe. It can be shown that an estimate of this displacement can be given by minimizing the following energy function (Fahmy et al., 2012):

$$E^t = E_{\mathrm{OF}}^t E_{\mathrm{cnt}}^t \quad (1)$$

The first term in (1) represents the optical flow constraint and is given by

$$E_{\mathrm{OF}}^t = \alpha \sum_{s=1}^{N} \left[u_s^t v_s^t 1 \right] \cdot J\left(\nabla I\right) \cdot \begin{bmatrix} u_s^t \\ v_s^t \\ 1 \end{bmatrix} \quad (2)$$

where α is a weighting parameter, $\nabla I = [I_x \; I_y \; I_t]^T$ and $J(\nabla I) = (\nabla I) \cdot (\nabla I)^T$.

The second term in (1) represents the desired properties of the myocardium contour, which can be formulated by the cost function proposed by (Kass et al., 1988), i.e.

$$E_{\mathrm{cnt}}^t = \sum_{s=1}^{N} \left(\beta \left| \frac{dP_s^t}{ds} \right|^2 + \gamma \left| \frac{d^2 P_s^t}{ds^2} \right|^2 \right) \quad (3)$$

where β and γ are weighting parameters.

To find the optimal solution $\left(u_s^t, v_s^t\right)^*$ that minimizes E^t in (1), an iterative greedy algorithm (Lam and Yan, 1994) was used. In this algorithm, the energy function is calculated at each pixel in the local neighbourhood of the contour point. Subsequently, the contour point is moved to the location with the minimum energy, and the process is repeated for all contour points until a convergence is reached.

The solution obtained by this method simultaneously minimizes both the OF and the contour properties constraints. This yields a solution that is more optimal than that of other methods that first calculate the displacement of the contour points using OF tracking before feeding it to an active snake algorithm as an initial contour (Mikić et al., 1998) or as an additional force term (Hamou and El-Sakka, 2010). This method required manual input of the contour on the first frame.

3.4. Block-matching algorithm (DS)

In this method, epicardial and endocardial contours were manually drawn on an initial timeframe, i.e., $t = 0$, and the contours were subsequently tracked by using the block-matching technique (Shi and Sun, 1999; Ourselin et al., 2000). The next point $P_s^{(t+1)}$ was estimated by a point Q along the radial line extending from the centroid of the contour to P_s^t. This point Q maximizes the following function

$$f = \text{Corr}\left(N_{P_s^t}, N_Q\right) e^{-\frac{|P_s^t - Q|}{(W_t/3)^2}} \quad (4)$$

where W_t is the average wall thickness calculated from the contours, N_P is 5×5 block around P and Corr(A, B) calculates the correlation between A and B (Song, 2011). This function was used to calculate contour points at timeframe $t + 1$.

A moving average 3 × 1 filter was applied to smooth the contour to remove any jumps in motion caused by erroneous estimation of motion parameters. Finally, the wall thickness W_{t+1} was recalculated from the resulting contours and the process was repeated for the next time frames. This method required manual input of the contours on the first frame.

3.5. Manually guide-point modeling assisted fitting of cardiac model (AU)

The Guide-Point Modeling technique (Li et al., 2010) was used to assist the fitting of a finite element cardiac model to the CMR data. This approach involves human observer input to refine the segmentation results by positioning a small number of guide points interactively on a sparse subset of slices and frames. Both long axis and short axis images were included in the analysis. The model incorporated the basal margin of the left ventricle as a plane, which was least squares fit to points placed by the user on the hinge points of the mitral valve in the long axis images. The model surfaces were influenced by the placement of user-defined guide points, and the automatic generation of edge points as well as the automated tracking of contours through all frames using non-rigid registration. The model was spatially and temporally consistent to reduce the amount of user interaction. However, inconsistency in breath-hold position can lead to mismatches between the short and long axis images. Images were manually shifted in-plane to compensate for breath-hold mis-registration, but individual slices may show errors in segmentation due to inconsistency with surrounding images in space and time. This expert-guided method has been previously validated in animals against autopsy LV mass, in patients with regional wall motion abnormalities against manually drawn contours, and in healthy volunteers against flow-derived measurements of cardiac output (Young et al., 2000). This method required expert approval of all slices and for all frames.

4. Consensus image estimation

The consensus images were estimated by using STA-PLE (Simultaneous Truth And Performance Level Estimation) method (Warfield et al., 2004). The method is essentially an instance of the EM algorithm (Dempster et al., 1977), which is a statistical method to estimate the maximum likelihood of missing or hidden data from an incomplete data set. In STAPLE, the incomplete data set is the collection of rater decisions, while the hidden data is the true segmentation image.

Let $\mathbf{D}$ be an $N \times R$ matrix of R rater decisions on N image pixels. Let $\mathbf{T}$ be an N-elements vector of the true segmentation image or the hidden data to be estimated. Each element of $\mathbf{D}$ and $\mathbf{T}$ contains a decision of one of L labels, i.e. $\mathbf{D}_{ij}, \mathbf{T}_i \in \{0, 1, \ldots, L - 1\}$. The complete data likelihood is defined as $L(\theta/\mathbf{D}, \mathbf{T}) = f(\mathbf{D}, \mathbf{T}/\theta)$. The parameter θ is defined for each rater.

Hence, there are θ_j parameters for $j = 1, \ldots, R$ and each contains $L \times L$ conditional probability values of a rater performance. In this application, $L = 2$ for myocardium (labeled as 1) and non-myocardium (labeled as 0) pixels. The conditional probability is therefore

$$\theta_j = \begin{bmatrix} p_j & (1-p_j) \\ (1-q_j) & q_j \end{bmatrix} \quad (5)$$

where

$$p_j = \Pr(\mathbf{D}_{ij}=1|\mathbf{T}_i=1) \quad (6)$$

$$q_j = \Pr(\mathbf{D}_{ij}=0|\mathbf{T}_i=0) \quad (7)$$

In (6), p_j basically defines the sensitivity or the probability that rater j correctly identifies myocardial pixels. Similarly in (7), q_j is the specificity or the probability that rater j correctly determines non-myocardial pixels.

The parameters θ are estimated by maximizing the log likelihood of the complete data, i.e.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \quad \ln \quad f(\mathbf{D}, \mathbf{T}|\theta). \quad (8)$$

Given an initial $\theta^{(0)}$ and the prior probability of the ground truth $f(\mathbf{T})$, the EM algorithm solves (8) by iterating between the expectation computation of $f(\mathbf{T}|\mathbf{D}, \theta^{(k)})$ (E-step), and estimating the parameters $\theta^{(k)}$ given the previous $\theta^{(k-1)}$ (M-step), where k denotes the iteration number (see more details in Warfield et al., 2004).

The STAPLE algorithm was initialized by calculating $\left(p_j^{(0)}, q_j^{(0)}\right)$ from the values of the expert-guided rater AU. The consensus was initialized to the AU rater since this was the one with the most manual input. This approach is stable enough to reach consistent optimal solutions. In the absence of raters with expert input, the majority vote rule from all contributing raters should give sufficient initial parameter values.

5. Setting the prior model

The STAPLE algorithm fundamentally combines the observed data (rater decisions) with a prior model, or $f(\mathbf{T})$, of the true segmentation image. A good prior model obviously leads to a better solution, but setting the prior model is not a trivial task.

The prior model can be defined spatially $\gamma_i = f(\mathbf{T}_i)$ or by a single global value $\gamma = f(\mathbf{T}_i)$, $\forall_i$ (Warfield et al., 2004). The spatially varying prior is recommended if a probabilistic atlas is available. If this is not available, a single global prior can be estimated from the sample mean of the relative proportion of each label, i.e.,

$$\gamma = f(\mathbf{T}_i=1) = \frac{1}{RN} \sum_{j=1}^{R} \sum_{i=1}^{N} D_{ij}.$$

(Zhu et al., 2008) argued that the prior model is not independent from the rater performance, i.e. $f(\mathbf{T}, p, q) \neq f(\mathbf{T})$. Hence, they proposed a Bayesian inference approach to model the prior in the presence of rater performance.

The single global prior (9) is suitable if the number of pixels for each label are comparable in the image space. However, for CMR images, the number of background pixels is much higher than the number of myocardial pixels (ranging from 50 up to more than 1,000 times). Applying γ on these images creates an adverse effect, particularly in the contentious areas where only a few raters decide them as myocardium. Figure 4 demonstrates this effect.

To avoid such a problem, a region of interest around the myocardium can be introduced to limit the number of pixels. In (Suinesiaputra et al., 2012), we used the expert-guided rater AU decisions to define the region of interest, as well as the global prior model, by setting the region size to include the same number of foreground and background pixels. Good results were obtained, but this solution did not treat all raters equally when deciding the prior.

If all raters have agreed upon a pixel, or

$$f\left(\mathbf{T}_i = l | \mathbf{D}_i, \theta\right) = 1 \quad \text{if} \quad \mathbf{D}_{ij} = l, \forall j = 1, \ldots, R, \quad (10)$$

then STAPLE will not change the pixel label during the iteration. It may therefore be beneficial to consider only pixels with rater disagreement during the STAPLE computation, which excludes all background and foreground pixels with agreement from all raters. This approach basically creates a region of interest defined by all raters uniformly. Consequently, the prior γ represents the proportion of undecided myocardial pixels in the input data.

In this study, we used the latter approach to determine the prior γ. The STAPLE implementation[2] supports this approach by enabling the “assign consensus voxels” (ACV) option. Based on our investigations, this approach returned similar results with our previous study (Suine-siaputra et al., 2012), which used the expert regions of interest to give the same number of foreground and background pixels (see Fig. 4).

6. Evaluation and objective test criteria

To avoid bias due to large background areas, we used regions of interest defined by the ACV option to calculate the accuracy and similarity indices. Let T_1 and T_0 be the number of pixels which characterise correctly myocardium and non-myocardium. Let F_1 and F_0 be the number of pixels, which misclassify myocardium and non-myocardium. The sensitivity (p), specificity (q), positive predictive value (PPV) and negative predictive value (NPV) are defined as follows

$$\begin{aligned} p = & \frac{T_1}{T_1 + F_0}, q = \frac{T_0}{T_0 + F_1} \\ \text{PPV} = & \frac{T_1}{T_1 + F_1}, \text{NPV} = \frac{T_0}{T_0 + F_0} \end{aligned} \quad (11)$$

To measure the similarity of two binary images, we used the Jaccard index, which measures the ratio of the overlap area. It is defined as

$$J\left(\mathscr{D}, \mathscr{T}\right) = \frac{|\mathscr{D} \bigcap \mathscr{T}|}{|\mathscr{D} \bigcup \mathscr{T}|}, \quad (12)$$

[2] http://crl.med.harvard.edu/software/STAPLE/index.php

where $/\cdot/$ denotes the cardinality of a set for $\mathscr{D}=1$ and $\mathscr{T}=1$. Note that we did not use the Dice index in this study, because the Jaccard indices are uniformly distributed, while the Dice indices are more inclined toward high values (Chang et al., 2009).

Clinical LV functional parameters were assessed by means of endocardial volume (EDV), epicardial volume (ESV), ejection fraction (EF) and LV mass. Endocardial and epicardial volumes were computed by using pixel summation method. The EF measures the volumetric fraction of blood pumped out of the LV, which is defined as EF = (EDV - ESV) /EDV × 100%. The LV mass was calculated by the myocardial volume at ED, i.e. subtraction of epicardial volume by endocardial volume, multiplied by the myocardial density (1.05 g/ml). These are standard clinical parameters used to clinically evaluate LV function from MRI (Salton et al., 2002).

In this paper, we also investigated an assessment test for the inclusion of an automated rater into the consensus. We propose a global measure based on the joint distribution of sensitivity (p) and specificity (q). Adapted from (Commowick and Warfield, 2010), the distributions of p and q are first transformed into normal-like distribution using Logit transformation (Collins et al., 1992):

$$\text{Logit}\,(X)=\ln\left(\frac{X}{1-X}\right), \quad (13)$$

where $X\in(0, 1)$ is a random variable. Note that (13) is undefined for $X = 0$ and $X = 1$. Since $0 \le p, q \le 1$, the Logit transformation (13) can be directly applied to the sensitivity and specificity distributions of each rater.

We used an unpaired two-sided t-test of the Bhattacharyya distances (Fukunaga, 1990) to test whether the distances between the automated rater performance distribution and those of the manual raters are different from the manual inter-rater distances. The Bhattacharrya distance for two normal rater distributions (r_1 and r_2) is given as follows

$$D_B\,(r_1,r_2)=\frac{1}{8}(\mu_{r_1}-\mu_{r_2})^T\left(\frac{\Sigma_{r_1}+\Sigma_{r_2}}{2}\right)^{-1}(\mu_{r_1}-\mu_{r_2})+\frac{1}{2}\ln\left(\frac{|\frac{1}{2}\,(\Sigma_{r_1}+\Sigma_{r_2})\,|}{\sqrt{|\Sigma_{r_1}||\Sigma_{r_2}|}}\right), \quad (14)$$

where $\mu\in\mathscr{R}^2$ is the estimated mean of a rater performance distribution, and Σ is the corresponding covariance matrix.

7. Experimental setup

All raters delineated myocardium in short-axis cine MRI, producing binary images with 0 (background) or 1 (myocardium) pixel labels. The STAPLE algorithm was applied on each 2D image slice independently. The STA-PLE algorithm estimated the consensus binary images with the ACV option enabled. Only images with more than two rater segmentations were included in the experiment. The outputs of the STAPLE algorithm were probability map images. A threshold value $p > 0.5$ was chosen empirically to determine myocardial pixels. We performed experiments on different threshold values between 0.2 to 0.7 and the average accuracies were similar within this range.

The consensus was initially applied to all five contributing raters, which we called CSALL. Another consensus (CSMAN) was built from three raters with manual intervention, i.e. AU, AO and DS. The two consensus sets were compared to investigate the effect of adding automated raters. Comparisons of each rater with CSMAN could then be performed to establish clinical functional parameters and performance distribution criteria for the incorporation of automated raters into the consensus.

In this paper, SCR and INR were used as examples of two automated raters which could possibly be incorporated into the baseline consensus (CSMAN) to improve the overall ground truth. Those raters passing the Bhattacharrya distances (14) t-test (within 95% confidence intervals of intra-manual rater distances) were included in a new consensus, called CS*.

8. Results

8.1. General rater performance

Figure 5 shows representative examples of segmentations from each rater with the two consensus. In general, STAPLE generated acceptable consensus images if the region of interest was limited to the rater disagreement area. The performance of each rater using both the CSMAN, CSALL and CS* as reference are shown in Table 2. The performance of the three manual raters (AU, AO and DS) were generally high for both sensitivity and specificity. For INR, it was high in sensitivity but low in specificity, while SCR was high in specificity but low in sensitivity. The highest similarity of segmented images were achieved by the manual raters. Automated raters achieved slightly lower similarity performances.

Clinical functional parameters are compared in Table 4. These show the average difference and standard deviation of the differences between each rater and the consensus, as well as the root mean squared error (RMSE). AU was the closest to the consensus. AO had a somewhat overestimated ESV relative to the other raters with manual input. Although INR EF bias was small, INR produced the largest deviation from the consensus mass and volumes. This was because INR included papillary muscles in the myocardium, not the blood pool, in contrast to the other raters (see Fig 5). This can also be seen in Table 4 where there are large mass biases for the INR rater.

Table 3 shows the average similarities between individual raters and between the three consensus segmentations (CSMAN, CSALL and CS*). Higher similarity was achieved between each rater and the consensus segmentations than between individual raters, which shows that the consensus segmentations act as a better reference ground truth than any one particular rater.

8.2. Regional and temporal variation of rater performance

Figure 6 compares PPV and NPV values in three regions of cardiac levels: base, mid-ventricle and apex. In general, PPV and NPV had slightly higher values and less scatter at the mid-ventricle, with reduced performance in apical slices.

Figure 7 shows the regional variation of agreement for a single representative case. Cohen's kappa coefficient (Cohen, 1960) was used to visualise areas of disagreement. The most disagreement was found at the apex and base, particularly around the outflow tract (arrow). Other areas of greater disagreement could be seen around the trabeculae and papillary muscles.

Figure 8 shows the temporal variation of the rater performance. PPV was used in this case, because it accounts for the accuracy of a rater to determine myocardium without the influence of the number of background pixels. Hence, the difficulty to segment thin walled myocardium compared to thick walled myocardium will be proportionally represented. The behaviour shown in Fig. 8 varied among raters, showing different methods had different temporal behaviour. Consistent with Table 4, AO performance was reduced at ES.

8.3. Assessment of automated raters for new consensus

Figure 9 shows a graphical representation of joint sensitivity and specificity distributions from all raters after being normalised with the Logit transformation (13). The four quadrants of this graph represent different types of segmentation results (good, poor, over and under segmentation). Ideally, a rater should have a distribution towards the top right corner of the figure, which are demonstrated by the manual raters (AU, AO and DS).

Table 5 shows the Bhattacharrya distances (14) between each rater distribution. The Bhattacharyya distances of INR vs. the manual raters were > 0.8, while those of SCR were < 0.3. The INR distances were significantly greater than the distances between manual raters ($p < 0.05$, CI = 95%), whereas the SCR distances were not significantly different from the distances between manual raters ($p = 0.49$). Thus the oversegmentation of INR seen in Fig. 9 was statistically significant, whereas the slight undersegmentation of SCR was not significantly different from the manual raters. On the basis of this test, the SCR rater was added with AO, AU and DS into CS*.

Table 2 also shows the rater performances against CS*. Performance results were quite consistent between the CSMAN, CSALL and CS* consensus segmentations. Clinical functional parameters (Table 4) show that the volume and mass differences of manual raters were maintained in CS* from CSMAN. The changes were all under 5 ml or g, which is clinically insignificant.

9. Discussion

STAPLE was designed to estimate hidden true segmentation images by maximizing the performance (sensitivity and specificity) of all raters simultaneously by using no knowledge other than their own decisions (Warfield et al., 2004). The resulting segmentation therefore represents the maximum possible performance of each rater when they need to reach a consensus amongst them.

9.1. The characteristics of input raters

The AU, AO and DS raters had manual input, either from manual contouring for initialisation (AO and DS) or by full spatiotemporal interaction (AU). For AO and DS, the initial contours placed on the first frame were tracked through the subsequent frames using an optical flow algorithm (for AO) and a block matching algorithm (for DS). The initial segmentation accuracy is therefore likely to have an effect on all frames. Since the manual input varied between raters, the PPV differences in Table 2 were therefore mainly due to the inter-observer variability.

Based on Table 2, the performance of the three manual raters were high, both in terms of identifying correct myocardium and labeling background pixels. The segmentation image similarities were also higher compared to automated raters. This high consistency was to be expected because of human expertise was involved during the segmentation process. To examine the performance of raters in subgroups of different cardiac function, we divided the cases into three groups: EF < 40%, 40% ≤ EF < 60%, and EF ≥ 60%. The results were similar in each case.

9.2. Locus of disagreement

Figure 7 shows a comparison of disagreements of a particular case between individual raters and between raters and the consensus. In this case, there is a large disagreement at base and apex, which corresponds to the difficulty to draw contours in these areas. The problem of where to determine the LV boundary at the outflow tract is still an open question which must be addressed by the community (see arrow in Fig. 7). Additional problem areas arise due to the low contrast ratio at apex, partly due to the partial volume effect, particularly at late-systole where the blood cavity is poorly seen.

Similarity indices from all raters were higher for midventricular slices, than apex and base slices. Thus raters did not experience a significant problem to segment midventricular myocardium, due to crisp endocardial and epicardial boundaries throughout the whole cardiac cycle. However, we found low similarity indices by INR with each consensus, because of the inclusion of papillary muscle in their algorithm (Fig. 5(b)). Figure 7 also shows this effect.

The temporal consistency of each rater to determine myocardium across normalised cardiac cycle was shown in Fig. 8. This was variable between raters, and indicated the range of each raters precision rate for segmenting myocardium. This can be useful for future development of segmentation algorithms. For example, one rater may wish to improve their algorithm on a specific timeframe, while others might wish to investigate improvement in all cardiac frames.

9.3. Adding automated raters to the consensus

Although raters with manual intervention (AO, AU and DS) consistently outperformed the automated raters (INR and SCR), there were benefits to adding automated raters to the consensus segmentation. The test for Bhattacharrya distances confirmed that the SCR rater could feasibly be included in the consensus, but that the INR rater should be excluded.

One benefit of adding automated raters is shown in Fig. 10 where the CSMAN consensus was improved by the addition of the SCR rater (CS*). In Fig. 11, we show three comparisons of the CS* consensus against the expert-guided rater AU, showing that the consensus can be more robust than an expert-guided rater. This was confirmed in Table 3, which showed that the agreement with the consensus was higher for all raters, indicating that the consensus is more robust than any individual rater.

Problem areas arise at the LV outflow tract (Fig. 10 bottom). This needs further consideration by the community before clinically meaningful segmentations can be derived. Other problems that are commonly encountered in the segmentation of myocardium are image artifacts due to gating or breathing issues, or off-resonance effects. The best way to reach a definite consensus for these issues is to have as many raters as possible.

The inclusion of automated methods is valuable since they have a deterministic response to image features. If there are only a limited number of expert manual raters available, automated methods can therefore add value to the consensus. In this way, we expect that the consensus estimate will become more robust as more participants contribute results.

9.4. Limitations

With only three raters in the initial consensus CSMAN, the test for inclusion of automated raters based on the Bhattacharrya distance lacks power. However, with more raters being contributed to the consensus, this test should become more powerful. Other tests are also possible. Our work demonstrates the ability of such tests in the application of consensus building and shows how raters can be compared and evaluated (in the example of the oversegmenting INR rater).

The STAPLE algorithm estimates the probability of each pixel independently. Hence, information about the neighboring pixel is not taken into account. As such, there is a possibility of forming an unwanted over-segmentation of the cavity inside the myocardium.

It would be beneficial to incorporate spatially correlated structure as suggested by (Warfield et al., 2004). Particularly for cardiac image segmentation, one might want to include spatiotemporal correlated structure embedded into the iteration to produce not only homogenous myocardial areas, but also smooth temporal myocardium across cardiac phases.

We found that STAPLE produced high quality consensus images if the region of interest was limited to the area of discrepancy between raters. Having a larger region of interest tended to produce over-segmentations (see Fig. 4). Good results were also obtained if the region of interest was limited to include roughly equal numbers of background and foreground pixels. This issue requires further investigation.

In this study, the consensus segmentation images were generated based on pixels. The drawback of this approach is that the myocardial borders are not smooth compared to continuous model-based or contour-based segmentation methods with sub-pixel resolution. Contour-based STA-PLE variation (Commowick and Warfield, 2009) may solve this problem, which will produce better and smoother myocardial borders.

10. Conclusion

We have presented results of generating consensus images for myocardium from a mixed of fully and semi automated segmentation algorithms. We have also investigated how adding automated raters can add value to the consensus segmentation. The results show that the STAPLE algorithm is a promising tool to generate the consensus images for cardiac MRI. Some improvements are still needed to be done, particularly in incorporating spatiotemporal information and resolving rater disagreements in some areas.

This collaborative framework is a first attempt towards establishing ground truth images for validation and benchmarking of segmentation algorithms. The number of raters involved in this study is still small. This is not a sufficient number of raters to produce a definite consensus. The more raters involved the better the ground truth images will become. Specific inclusion criteria for automated raters have been introduced to maintain the quality of the consensus, which are based on the distribution of rater performance.

Currently, we are continuing this collaboration study as an open ongoing project through the Cardiac Atlas Project website. The consensus images will be made available to research groups who contribute to this work, and to other researchers for purposes other than segmentation. In providing these consensus images, we require that they are not used to train new segmentation algorithms, so they can continue to be used as an independent validation. The consensus images will also be updated as new segmentation results are contributed.

Acknowledgments

The Cardiac Atlas Project was supported by National Heart, Lung and Blood Institute, USA (R01HL087773). A.S. Fahmy was supported in part by a National Science Foundation and Science Technology Development Fund (NSF-STDF) grant number 1979. J. Margeta was supported by Microsoft Research through its PhD Scholarship Programme and part of his work was funded by the European Research Council through the ERC Advanced Grant MedYMA (on Biophysical Modeling and Analysis of Dynamic Medical Images). D.C. Lee and A.H. Kadish are DETERMINE investigators. The DETERMINE Trial was supported by St Jude Medical, Inc.; and the National Heart, Lung and Blood Institute (R01HL91069). A list of participating DETERMINE investigators can be found at http://www.clinicaltrials.gov.

References

Archip N, Jolesz FA, Warfield SK. A validation framework for brain tumor segmentation. Acad Radiol. 2007; 14:1242–51. [PubMed: 17889341]

Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, Jacob DR Jr, Kronmal R, Liu K, Nelson JC, O'Leary D, Saad MF, Shea S, Szklo M, Tracy RP. Multi-ethnic study of atherosclerosis: objectives and design. Am J Epidemiol. 2002; 156:871–81. [PubMed: 12397006]

Chang HH, Zhuang AH, Valentino DJ, Chu WC. Performance measure characterization for evaluating neuroimage segmentation algorithms. Neuroimage. 2009; 47:122–35. [PubMed: 19345740]

Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960; 20:37–46.

Collins J, Mancilulli M, Hohlfeld R, Finch D, Sandri G, Shtatland E. A random number generator based on the logit transform of the logistic variable. Computers in Physics. 1992; 6:630–632.

Commowick O, Warfield SK. A continuous STAPLE for scalar, vector, and tensor images: an application to DTI analysis. IEEE Trans Med Imaging. 2009; 28:838–46. [PubMed: 19272988]

Commowick O, Warfield SK. Estimation of inferential uncertainty in assessing expert segmentation performance from STA-PLE. IEEE Trans Med Imaging. 2010; 29:771–80. [PubMed: 20199913]

Criminisi A, Shotton J, Konukoglu E. Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. Foundations and Trends in Computer Graphics and Vision. 2012; 7:81–227.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via EM algorithm. J Roy Stat Soc B Met. 1977; 39:1–38.

Fahmy, A.; Al-Agamy, A.; Khalifa, A. Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges. Springer: 2012. Myocardial Segmentation Using Contour-Constrained Optical Flow Tracking; p. 120-128.

Fonseca CG, Backhaus M, Bluemke DA, Britten RD, Chung JD, Cowan BR, Dinov ID, Finn JP, Hunter PJ, Kadish AH, Lee DC, Lima JAC, Medrano-Gracia P, Shivkumar K, Suinesiaputra A, Tao W, Young AA. The Cardiac Atlas Project – An imaging database for computational modeling and statistical atlases of the heart. Bioinformatics. 2011; 27:2288–2295. [PubMed: 21737439]

Frangi AF, Niessen WJ, Viergever MA. Three-dimensional modeling for functional analysis of cardiac images: a review. IEEE Trans Med Imaging. 2001; 20:2–25. [PubMed: 11293688]

Fukunaga, K. Introduction to Statistical Pattern Recognition. 2nd edition. Academic Press; 1990.

Geremia E, Clatz O, Menze BH, Konukoglu E, Criminisi A, Ayache N. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. Neuroimage. 2011; 57:378–90. [PubMed: 21497655]

van Ginneken B, Armato SG 3rd, de Hoop B, van Amelsvoort-van de Vorst S, Duindam T, Niemeijer M, Murphy K, Schilham A, Retico A, Fantacci ME, Camarlinghi N, Bagagli F, Gori I, Hara T, Fujita H, Gargano G, Bellotti R, Tangaro S, Bolaños L, De Carlo F, Cerello P, Cristian Cheran S, Lopez Torres E, Prokop M. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study. Med Image Anal. 2010; 14:707–22. [PubMed: 20573538]

Guetter, C.; Xue, H.; Chefd'hotel, C.; Guehring, J. Efficient symmetric and inverse-consistent deformable registration through interleaved optimization. Biomedical Imaging: From Nano to Macro; IEEE International Symposium on; 2011. 2011. p. 590-593.

Hameeteman K, Zuluaga MA, Freiman M, Joskowicz L, Cuisenaire O, Valencia LF, Gülsün MA, Krissian K, Mille J, Wong WCK, Orkisz M, Tek H, Hoyos MH, Benmansour F, Chung ACS, Rozie S, van Gils M, van den Borne L, Sosna J, Berman P, Cohen N, Douek PC, Sánchez I, Aissat M, Schaap M, Metz CT, Krestin GP, van der Lugt A, Niessen WJ, van Walsum T. Evaluation framework for carotid bifurcation lumen segmentation and stenosis grading. Med Image Anal. 2011; 15:477–88. [PubMed: 21419689]

Hamou AK, El-Sakka MR. Optical flow active contours with primitive shape priors for echocardiography. EURASIP J. Adv. Signal Process. 2010; 2010; 9(2-9):2.

Hoyte L, Ye W, Brubaker L, Fielding JR, Lockhart ME, Heilbrun ME, Brown MB, Warfield SK, Pelvic Floor Disorders Network. Segmentations of MRI images of the female pelvic floor: a study of interand intra-reader reliability. J Magn Reson Imaging. 2011; 33:684–91. [PubMed: 21563253]

Hwee J, Louie AV, Gaede S, Bauman G, D'Souza D, Sexton T, Lock M, Ahmad B, Rodrigues G. Technology assessment of automated atlas based segmentation in prostate bed contouring. Radiat Oncol. 2011; 6:110. [PubMed: 21906279]

Ishida M, Kato S, Sakuma H. Cardiac MRI in ischemic heart disease. Circ J. 2009; 73:1577–88. [PubMed: 19667487]

Jolly, MP. Proceedings of the 11th international conference on Medical Image Computing and Computer-Assisted Intervention - Part I. Springer-Verlag; Berlin, Heidelberg: 2008. Automatic recovery of the left ventricular blood pool in cardiac cine MR images; p. 110-118.

Jolly, MP.; Guetter, C.; Guehring, J. Cardiac segmentation in MR cine data using inverse consistent deformable registration. Biomedical Imaging: From Nano to Macro; IEEE International Symposium on; 2010. 2010. p. 484-487.

Jolly, MP.; Guetter, C.; Lu, X.; Xue, H.; Guehring, J. Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges. Springer; 2012. Automatic Segmentation of the Myocardium in Cine MR Images Using Deformable Registration; p. 98-108.

Kadish AH, Bello D, Finn JP, Bonow RO, Schaechter A, Subacius H, Albert C, Daubert JP, Fonseca CG, Goldberger JJ. Rationale and design for the Defibrillators to Reduce Risk by Magnetic Resonance Imaging Evaluation (DE-TERMINE) trial. J Cardiovasc Electrophysiol. 2009; 20:982–7. [PubMed: 19493153]

Karamitsos TD, Myerson SG. The role of cardiovascular magnetic resonance in the evaluation of valve disease. Prog Cardiovasc Dis. 2011; 54:276–86. [PubMed: 22014494]

Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models. Int J Comput Vis. 1988; 1:321–331.

Lam KM, Yan H. Fast greedy algorithm for active contours. Electronics Letters. 1994; 30:21–2.

Lempitsky, VS.; Verhoek, M.; Noble, JA.; Blake, A. FIMH. Springer; 2009. Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography; p. 447-456.

Li B, Liu Y, Occleshaw CJ, Cowan BR, Young AA. In-line automated tracking for ventricular function with magnetic resonance imaging. JACC Cardiovasc Imaging. 2010; 3:860–6. [PubMed: 20705268]

Liu T, Li H, Wong K, Tarokh A, Guo L, Wong STC. Brain tissue segmentation based on DTI data. Neuroimage. 2007; 38:114–23. [PubMed: 17804258]

Lo P, van Ginneken B, Reinhardt J, de Bruijne M. Extraction of Airways From CT (EXACT'09). IEEE Trans Med Imaging. 2012; 31:2093–2107. [PubMed: 22855226]

Lu, X.; Georgescu, B.; Jolly, MP.; Guehring, J.; Young, A.; Cowan, B.; Littmann, A.; Comaniciu, D. Proceedings of the 13th international conference on Medical image computing and computer-assisted intervention: Part I. Springer-Verlag; Berlin, Heidelberg.: 2010. Cardiac anchoring in MRI through context modeling; p. 383-390.

Margeta, J.; Geremia, E.; Criminisi, A.; Ayache, N. Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges. Springer; 2012. Layered Spatio-temporal Forests for Left Ventricle Segmentation from 4D Cardiac MRI Data; p. 109-119.

Mikić I, Krucinski S, Thomas JD. Segmentation and tracking in echocardiographic sequences: active contours guided by optical flow estimates. IEEE Trans Med Imaging. 1998; 17:274–84. [PubMed: 9688159]

Murphy K, van Ginneken B, Reinhardt JM, et al. Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge. IEEE Trans Med Imaging. 2011; 30:1901–20. [PubMed: 21632295]

Muzzarelli S, Ordovas K, Higgins CB. Cardiovascular MRI for the assessment of heart failure: focus on clinical management and prognosis. J Magn Reson Imaging. 2011; 33:275–86. [PubMed: 21274968]

Nyúl LG, Udupa JK. On standardizing the MR image intensity scale. Magn Reson Med. 1999; 42:1072–81. [PubMed: 10571928]

Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. Pharmacogenomics. 2005; 6:639–46. [PubMed: 16143003]

Ourselin, S.; Roche, A.; Prima, S.; Ayache, N. Block Matching: A general framework to improve robustness of rigid registration of medical images. In: Delp, S.; DiGoia, A.; Jaramaz, B., editors.

Medical Image Computing and Computer-Assisted Intervention – MICCAI. Springer; Berlin/ Heidelberg: 2000. 2000. p. 557-66.

Paetsch I, Jahnke C, Ferrari VA, Rademakers FE, Pellikka PA, Hundley WG, Poldermans D, Bax JJ, Wegscheider K, Fleck E, Nagel E. Determination of interobserver variability for identifying inducible left ventricular wall motion abnormalities during dobutamine stress magnetic resonance imaging. Eur Heart J. 2006; 27:1459–1464. [PubMed: 16613929]

Pennell DJ. Cardiovascular magnetic resonance. Circulation. 2010; 121:692–705. [PubMed: 20142462]

Petitjean C, Dacher JN. A review of segmentation methods in short axis cardiac MR images. Med Image Anal. 2011; 15:169–84. [PubMed: 21216179]

Radau P, Lu Y, Connelly K, Paul G, Dick AJ, Wright GA. Evaluation Framework for Algorithms Segmenting Short Axis Cardiac MRI. The MIDAS Journal - Cardiac MR Left Ventricle Segmentation Challenge. 2009:49.

Salton CJ, Chuang ML, O'Donnell CJ, Kupka MJ, Larson MG, Kissinger KV, Edelman RR, Levy D, Manning WJ. Gender differences and normal left ventricular anatomy in an adult population free of hypertension. A cardiovascular magnetic resonance study of the Framingham Heart Study Offspring cohort. J Am Coll Cardiol. 2002; 39:1055–60. [PubMed: 11897450]

Schaap M, Metz CT, van Walsum T, van der Giessen AG, Weustink AC, Mollet NR, Bauer C, Bogunović H, Castro C, Deng X, Dikici E, O'Donnell T, Frenay M, Friman O, Hernández Hoyos M, Kitslaar PH, Krissian K, Kühnel C, Luengo-Oroz MA, Orkisz M, Smedby O, Styner M, Szymczak A, Tek H, Wang C, Warfield SK, Zambal S, Zhang Y, Krestin GP, Niessen WJ. Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms. Med Image Anal. 2009; 13:701–14. [PubMed: 19632885]

Shattuck DW, Prasad G, Mirza M, Narr KL, Toga AW. Online resource for validation of brain segmentation methods. Neuroimage. 2009; 45:431–9. [PubMed: 19073267]

Shi, YQ.; Sun, H. Image and Video Compression for Multimedia Engineering. 1st edition. CRC Press, Inc.; Boca Raton, FL, USA: 1999.

Song BC. A fast normalized cross correlation-based block matching algorithm using multilevel Cauchy-Schwartz inequality. ETRI Journal. 2011; 33:401–406.

Suinesiaputra, A.; Cowan, B.; Finn, J.; Fonseca, C.; Kadish, A.; Lee, D.; Medrano-Gracia, P.; Warfield, S.; Tao, W.; Young, A. Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges. Springer; 2012. Left Ventricular Segmentation Challenge from Cardiac MRI: A Collation Study; p. 88-97.

Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging. 2004; 23:903–21. [PubMed: 15250643]

Xing F, Soleimanifard S, Prince JL, Landman BA. Statistical fusion of continuous labels: Identification of cardiac landmarks. Proc Soc Photo Opt Instrum Eng. 2011:7962.

Young AA, Cowan BR, Thrupp SF, Hedley WJ, Dell'Italia LJ. Left ventricular mass and volume: fast calculation with guide-point modeling on MR images. Radiology. 2000; 216:597–602. [PubMed: 10924592]

Zhu Y, Huang X, Wang W, Lopresti D, Long R, Antani S, Xue Z, Thoma G. Balancing the role of priors in multi-observer segmentation evaluation. J Signal Process Syst. 2008; 55:185–207. [PubMed: 20523759]

- A collaboration to establish consensus ground truth segmentations for cardiac MRI.
- Consensus contours combined automated algorithms with expert rater by using STAPLE.
- Limiting regions of rater discrepancy maintained high quality consensus images.
- More raters would further increase the quality of consensus segmentation images.
- An objective measure based on global rater performance is introduced to test a candidate rater.

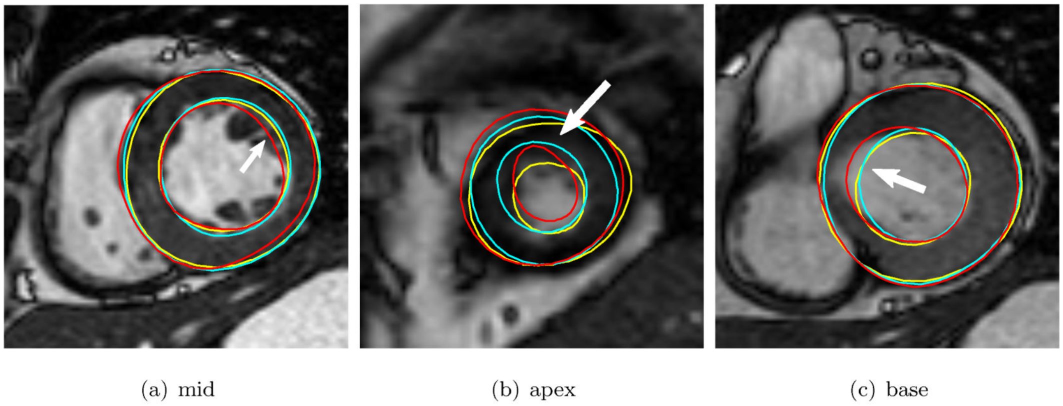


Figure 1.

Examples of expert rater differences (white arrows) to delineate myocardium due to different interpretation of where the contours should lie.

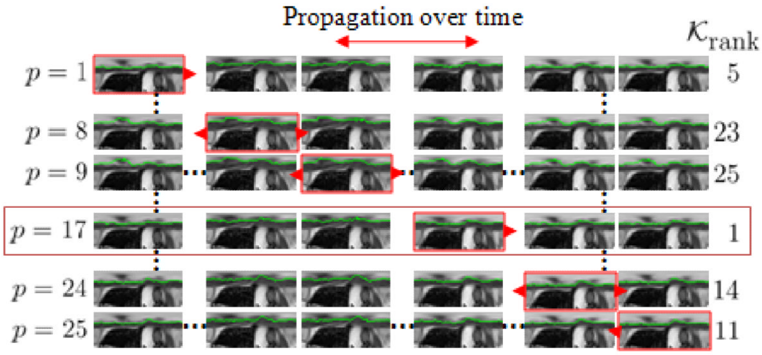


Figure 2.

Slice segmentation in SCR algorithm: for each frame *p* = 1, …, *P*, recover a contour using Dijkstra’s algorithm in polar space, propagate the contour to all other frames, and repeat. Choose the combination of contours (in this case recovered from frame 17) with the lowest cost.

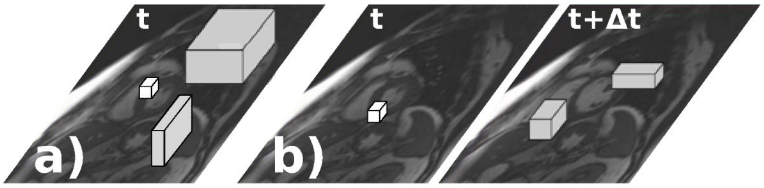


Figure 3.

Illustration of extracted image based features in layered spatio-temporal forest algorithm. a) Context rich features (Geremia et al., 2011) measuring differences between source regional average intensity centered at tested voxel (white box) and the sum of remote region averages (gray boxes). b) Extension of the features to the spatio-temporal domain, where remote regions are placed in a frame offset from the current *t* by a constant value Δ *t*.

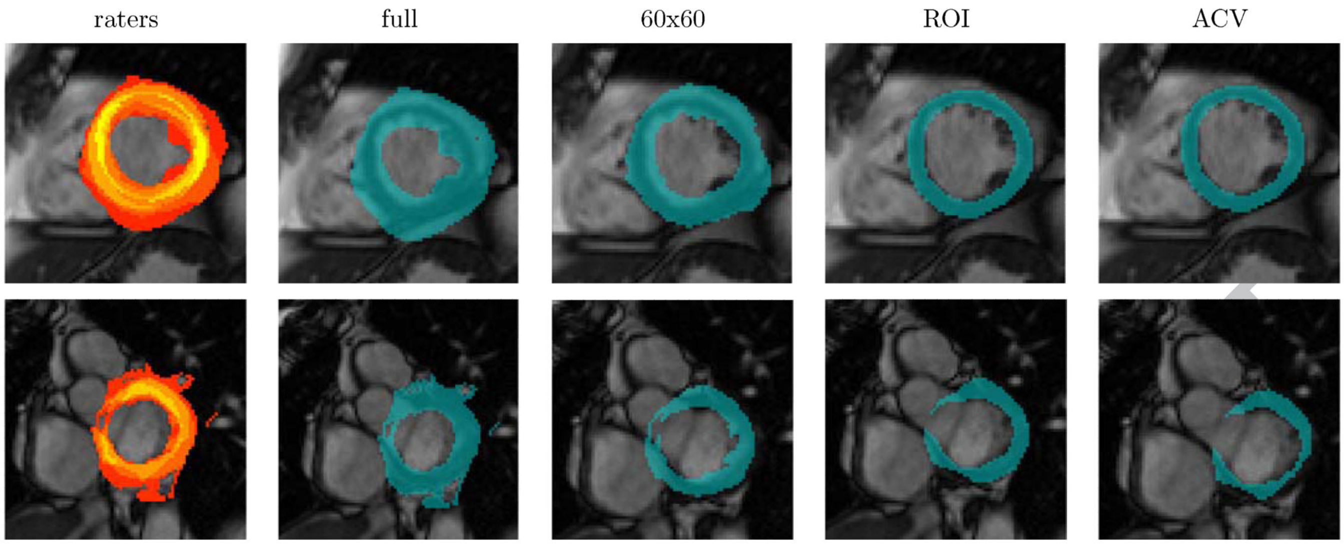
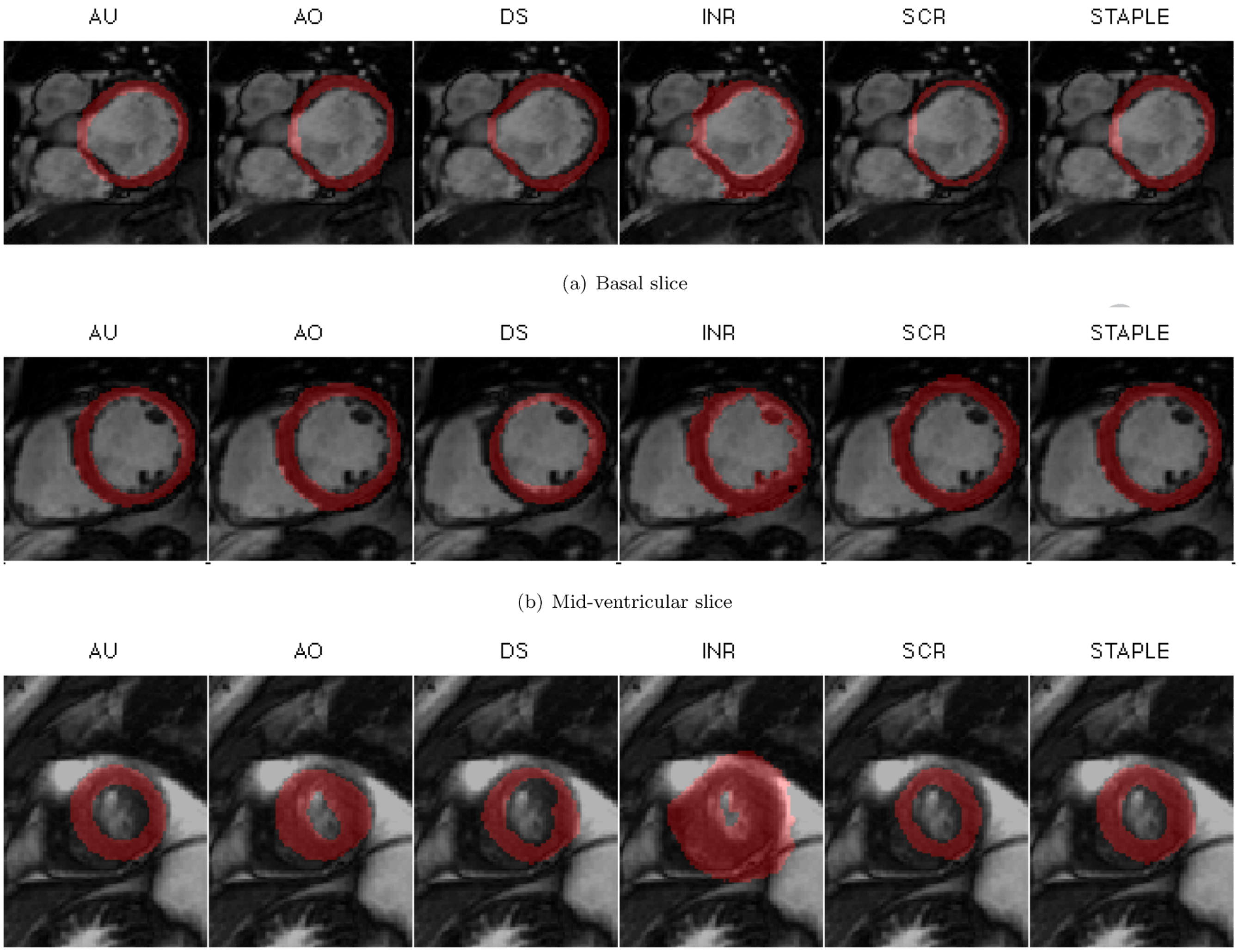


Figure 4.
The effects of different global priors to the STAPLE results on two cases. The leftmost figures are rater decisions on top of each other with color scale from red = 1 rater to yellow = 5 raters. The remaining four right figures are STAPLE computations with different ways to compute global prior values. The ‘full’ column figures are STAPLE images from the whole image region, the ‘60×60’ column shows STAPLE images from cropped rater images by 60 × 60 rectangular pixels, the ROI denotes region of interest approach around myocardium defined by the expert-guided rater (AU) to give the same number of foreground and background pixels, and the ACV stands for “assign consensus voxels” option where STAPLE only included pixels with rater disagreements during the estimation.



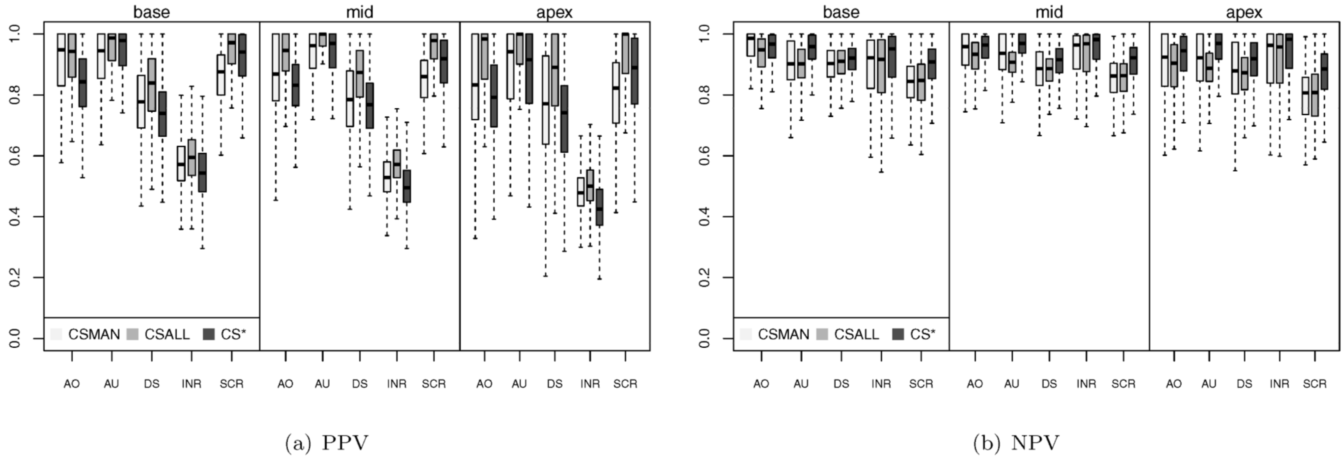
(a) Basal slice

(b) Mid-ventricular slice

(c) Apical slice

Figure 5.

Three representative examples of raters and CS* taken from base (a), mid-ventricular (b) and apical (c) slices.



(a) PPV

(b) NPV

Figure 6.

Comparison of PPV and NPV values from each rater between CSMAN and CSALL consensus.

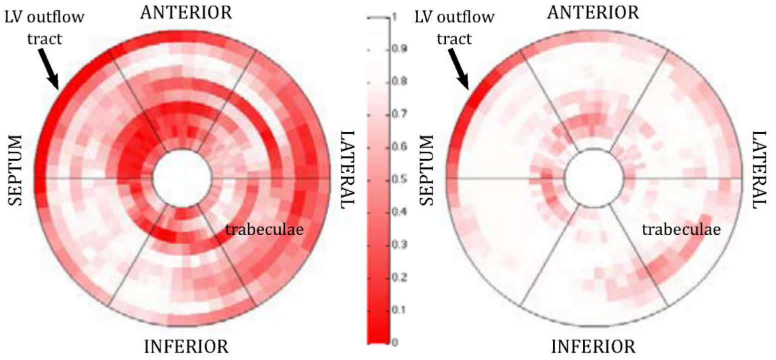


Figure 7.

Bullseye plots of rater disagreement from a particular case. Left: rater-rater disagreements. Right: rater-consensus disagreement. Color coded values are median of weighted Cohen’s kappa coefficients ranging from poor (red) to good (white) agreement.

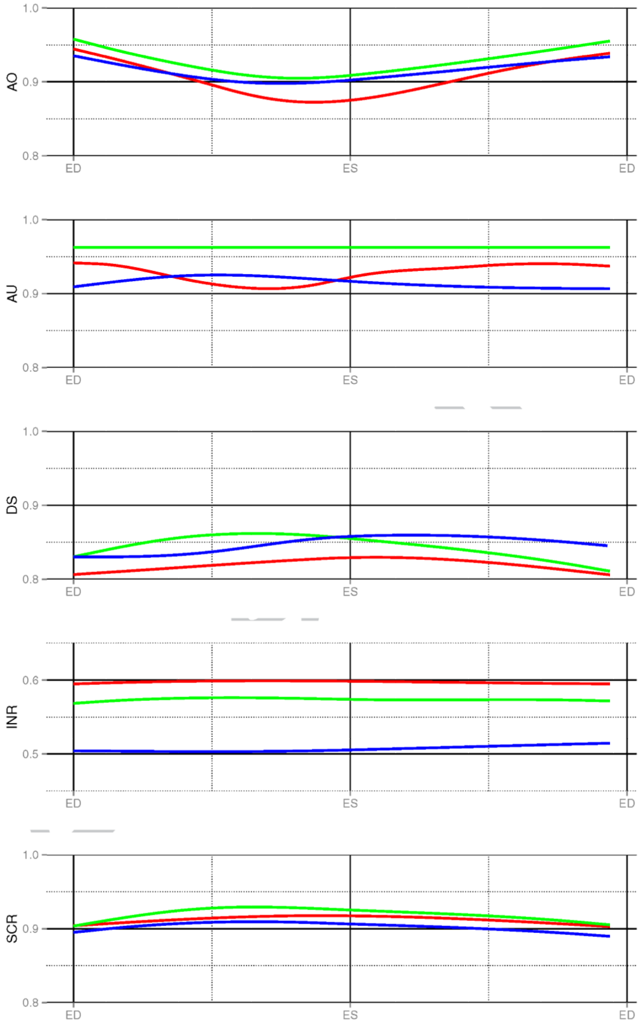


Figure 8.

Average positive prediction values (precision rates) over cardiac frames (ED=end-diastole, ES=end-systole). Colors indicate short-axis levels: red = basal slices, green = mid-slices, blue = apical slices.

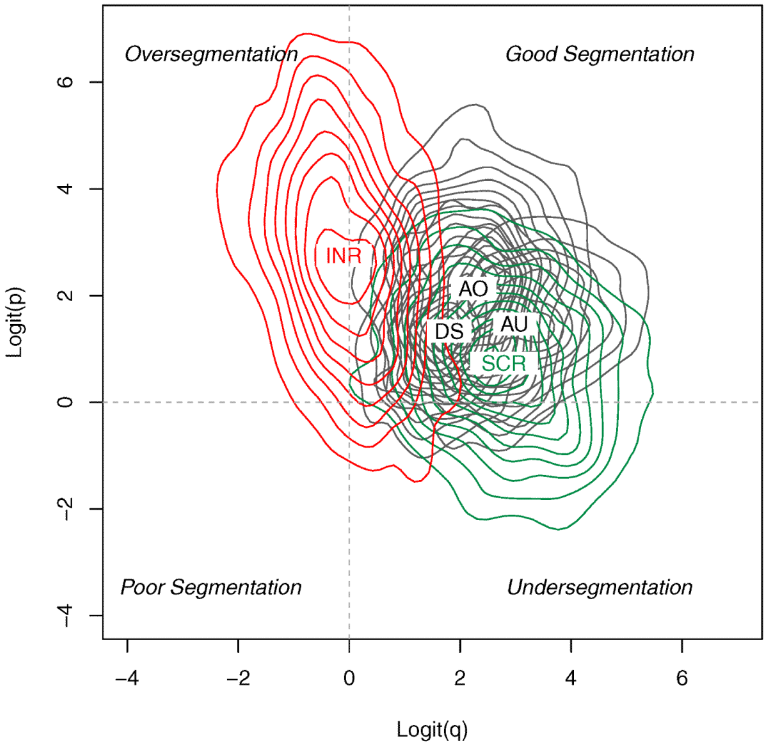


Figure 9.

Distributions of sensitivity and specificity values after Logit transformation (13). The value of Logit(0) is equal to sensitivity/specificity at 0.5. Therefore regions centered at the origin (dashed lines) define four characteristics of segmentation results. The rater labels are at the peak of each distribution, with the colors are: INR in red, SCR in green, and manual raters (AU, AO and DS) in black.

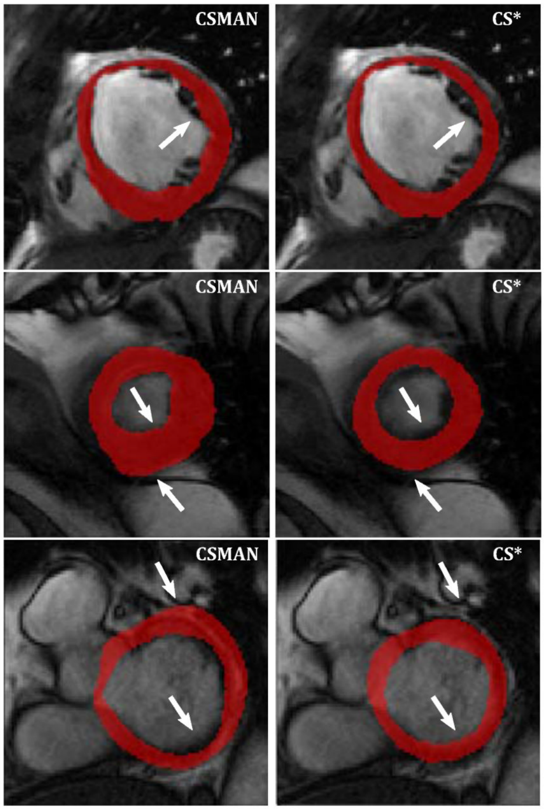


Figure 10.
Some comparisons between CSMAN (left) and CS* (right) consensus images, which show that adding automated raters can benefit the consensus in some cases.

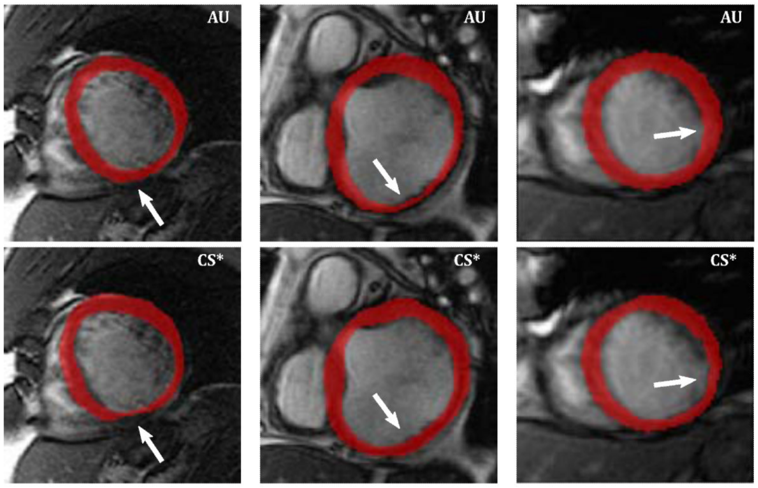


Figure 11.

Some comparisons between AU (top) with the CS* consensus (bottom). These examples show that the consensus can be better than an individual expert-guided rater.

Table 1

Patient data characteristics. Numerical data are expressed by ‘average (min–max)’ values.

Sex		
Male:	73 (76.8%)	
Female:	22 (23.2%)	
Age:	62.73	(34–84)
Systolic blood presure (mmHg):	122.96	(70–195)
Diastolic blood presure (mmHg):	71.49	(42–106)
Heart rate (bpm):	67.41	(45–105)

Table 2

Rater performance comparisons between **CSMAN**, **CSALL** and **CS***. All values are expressed as 'mean (standard deviation)'.

		sensitivity	specificity	PPV	NPV	Jaccard
AO	CSMAN	0.88 (0.15)	0.92 (0.09)	0.87 (0.13)	0.93 (0.08)	0.79 (0.19)
	CSALL	0.85 (0.15)	0.95 (0.07)	0.92 (0.11)	0.92 (0.07)	0.79 (0.16)
	CS*	0.88 (0.15)	0.91 (0.06)	0.82 (0.12)	0.94 (0.06)	0.74 (0.16)
AU	CSMAN	0.85 (0.16)	0.95 (0.07)	0.91 (0.13)	0.92 (0.08)	0.80 (0.19)
	CSALL	0.80 (0.15)	0.97 (0.08)	0.94 (0.12)	0.90 (0.07)	0.77 (0.17)
	CS*	0.89 (0.13)	0.96 (0.06)	0.91 (0.13)	0.95 (0.06)	0.84 (0.17)
DS	CSMAN	0.80 (0.18)	0.86 (0.10)	0.77 (0.17)	0.88 (0.10)	0.67 (0.21)
	CSALL	0.78 (0.16)	0.90 (0.09)	0.84 (0.15)	0.88 (0.08)	0.69 (0.17)
	CS*	0.80 (0.17)	0.86 (0.08)	0.74 (0.15)	0.90 (0.08)	0.64 (0.18)
INR	CSMAN	0.88 (0.17)	0.53 (0.18)	0.54 (0.09)	0.91 (0.10)	0.49 (0.10)
	CSALL	0.86 (0.20)	0.60 (0.16)	0.56 (0.11)	0.90 (0.11)	0.51 (0.13)
	CS*	0.89 (0.17)	0.56 (0.15)	0.50 (0.10)	0.93 (0.09)	0.43 (0.10)
SCR	CSMAN	0.66 (0.22)	0.92 (0.06)	0.83 (0.14)	0.83 (0.08)	0.59 (0.19)
	CSALL	0.63 (0.24)	0.98 (0.04)	0.91 (0.16)	0.83 (0.09)	0.61 (0.24)
	CS*	0.74 (0.23)	0.96 (0.05)	0.87 (0.16)	0.89 (0.09)	0.69 (0.23)

Table 3

The average of Jaccard similarity indices between individual raters and between consensus images for comparison.

	AO	**AU**	**DS**	**INR**	**SCR**
AO	—	0.64	0.59	0.44	0.53
AU	0.64	—	0.57	0.40	0.56
DS	0.59	0.57	—	0.42	0.48
INR	0.44	0.40	0.42	—	0.34
SCR	0.53	0.56	0.48	0.34	—
CSMAN	0.79	0.80	0.67	0.49	0.59
CSALL	0.79	0.77	0.69	0.51	0.61
CS*	0.74	0.84	0.64	0.43	0.69

Table 4

Clinical LV function differences between each rater from the **CSMAN**, **CSALL** and **CS*** consensus. All values are expressed in ‘mean (standard deviation) [RMSE = root of mean squared error]’.

Rater		CSMAN			CSALL			CS*		
AO	EDV (ml)	0.37	(8.49)	[8.46]	−2.37	(16.65)	[16.65]	−0.70	(8.78)	[8.76]
AU		−1.80	(7.38)	[7.56]	3.87	(8.81)	[9.53]	−2.90	(7.05)	[7.58]
DS		4.14	(14.39)	[14.90]	−6.43	(29.67)	[29.88]	2.75	(13.17)	[13.38]
INR		−55.94	(40.78)	[69.10]	−73.32	(47.74)	[86.90]	−55.65	(41.03)	[69.00]
SCR		19.02	(29.10)	[34.63]	1.85	(30.95)	[30.68]	17.79	(17.78)	[25.08]
AO	ESV (ml)	−11.89	(16.33)	[20.13]	−12.69	(24.67)	[27.49]	−13.50	(19.54)	[23.66]
AU		−2.23	(11.69)	[11.84]	1.40	(14.35)	[14.26]	−4.24	(5.85)	[7.20]
DS		11.75	(12.41)	[17.04]	7.97	(16.34)	[17.91]	9.21	(8.82)	[12.71]
INR		−38.43	(31.55)	[49.62]	−48.11	(33.21)	[58.06]	−39.76	(27.74)	[48.39]
SCR		15.88	(27.35)	[31.50]	6.47	(23.49)	[24.12]	14.28	(15.73)	[21.18]
AO	EF (%)	9.58	(10.67)	[14.30]	7.79	(13.76)	[15.67]	10.51	(11.97)	[15.88]
AU		0.60	(7.63)	[7.62]	0.94	(6.87)	[6.86]	1.74	(4.89)	[5.16]
DS		−7.98	(9.86)	[12.64]	−9.95	(7.92)	[12.56]	−6.55	(7.18)	[9.69]
INR		−1.33	(24.10)	[24.01]	−2.45	(22.42)	[22.32]	1.18	(22.01)	[21.92]
SCR		−2.63	(9.76)	[10.06]	−3.80	(7.43)	[8.27]	−1.42	(6.45)	[6.57]
AO	LV mass (g)	−1.81	(6.42)	[6.64]	−24.69	(16.86)	[29.70]	0.38	(7.24)	[7.21]
AU		−2.14	(6.71)	[7.01]	−15.41	(13.14)	[20.10]	−0.36	(7.24)	[7.21]
DS		6.76	(10.32)	[12.29]	−21.47	(20.46)	[29.28]	8.69	(11.72)	[14.53]
INR		124.35	(36.08)	[129.42]	90.23	(32.52)	[95.35]	127.80	(34.32)	[132.28]
SCR		−1.99	(25.67)	[25.61]	−36.18	(26.26)	[44.39]	0.75	(19.89)	[19.79]

Table 5

Bhattacharyya distances between each distributions in Fig. 9.

	AU	DS	SCR	INR
AO	0.19	0.13	0.18	0.89
AU	—	0.26	0.14	1.32
DS		—	0.26	0.84
SCR			—	1.06