

Left Ventricular Segmentation Challenge From Cardiac MRI: A Collation Study

Avan Suinesiaputra¹, Brett R. Cowan¹, J. Paul Finn², Carissa G. Fonseca², Alan H. Kadish³, Daniel C. Lee³, Pau Medrano-Gracia¹, Simon K. Warfield⁴, Wenchao Tao², and Alistair A. Young¹

¹ Auckland Bioengineering Institute, University of Auckland, New Zealand

² Department of Radiological Sciences, University of California Los Angeles, USA

³ Division of Cardiology, Northwestern University, USA

⁴ Computational Radiology Laboratory, Harvard Medical School, USA

Abstract. This paper presents collated results from the left ventricular (LV) cardiac MRI segmentation challenge as part of STACOM'11. Clinical cases from patients with myocardial infarction (100 test and 100 validation cases) were randomly selected from the Cardiac Atlas Project (CAP) database. Two independent sets of expert (manual) segmentation from different sources that are available from the CAP database were included in this study. Automated segmentations from five groups were contributed in the challenge. The total number of cases with segmentations from all seven raters was 18. For these cases, a ground truth “consensus” segmentation was estimated based on all raters using an Expectation-Maximization (EM) method (the STAPLE algorithm).

1 Introduction

In cardiac MRI, the LV segmentation is typically performed to derive important clinical indices such as LV mass and volume. The current clinical standard is manual contouring of the myocardial boundaries, a time consuming and error-prone process, requiring substantial training. The development of automated segmentation algorithms has been problematic due to the lack of “ground truth” in real clinical cases. Even expert manually drawn segmentations still suffer from inter- and intraobserver variability. This problem particularly applies in cardiac imaging, where the presence of papillary muscles, the heart dynamics, and soft tissue contrast variations are just some of the problematic areas in cardiac MRI.

In this segmentation challenge, we created a framework to solve this problem by providing the same data set to researchers to test their segmentation algorithms and also to estimate better set of ground truth segmentations at the same time. We applied the EM-based STAPLE method [6] to estimate the consensus ground truth segmentations. Therefore, the challenge was performed as a collaborative work rather than a competition. A large data set of clinical cardiac MRI cases was made available through the Cardiac Atlas Project⁵ [2]. By using the

⁵ <http://www.cardiacatlas.org>

Table 1. Baseline characteristics of the data used in this challenge.

	test set (N=100)	validation set (N=100)
EDV (ml)	193.86 (46.45)	199.44 (54.97)
ESV (ml)	113.41 (43.44)	123.80 (53.81)
LV mass (gr)	172.24 (42.57)	165.38 (40.30)
EF (%)	42.95 (10.88)	39.87 (11.25)
SV (ml)	80.41 (18.65)	75.59 (18.62)

EDV = endocardial volume at ED, ESV = endocadial volume at ES, EF = ejection fraction, and SV = stroke volume.

same data set, confounding difficulties to compare segmentation results between peers can therefore be eliminated.

2 Methods

2.1 Cardiac MRI Data

Cardiac MR images were randomly selected from the DETERMINE (Defibrillators To Reduce Risk by Magnetic Resonance Imaging Evaluation) cohort [4]. This study consists of patients with coronary artery disease and prior myocardial infarction. Two separate groups were defined as test (N=100) and validation (N=100) groups, by random selections (see Table 1). Cine MR images in short-axis and long-axis views were selected for this challenge. These MR images were acquired by using a Steady-State Free Precession (SSFP) pulse sequence. MRI parameters varied between cases, giving a heterogenous mix of scanner types and imaging parameters consistent with typical clinical cases.

2.2 Raters

Five automated raters (SCR, INR, DS, AO and EM) and two expert raters (AU and NU) participated in this study. Rater descriptions are given in Table 2. Two raters (SCR and INR) were fully automatic, although SCR required repositioning the center of LV segmentation in four cases. Three raters required some manual interactions, either by drawing initial contours (DS, EM and AO) or by having some parameter initialization (EM). One rater (INR) used the test dataset to train the algorithm, the others did not.

The expert NU rater was a manually drawn myocardial contour, traced by the DETERMINE MRI core laboratory using QMass software (Medis, Leiden, the Netherlands), while the AU rater was an expert-guided interactive customization of a finite element heart model using Cardiac Image Modeller (CIM) software (AMRG, Auckland, New Zealand). To generate the intersection between cardiac MRI with the 3D AU models and the image planes, the CAPClient software was used⁶.

⁶ The CAPClient is an open source software, available for download at <http://www.cardiacatlas.org/web/guest/tools>.

Table 2. Rater characteristics.

Rater	Method description	Dimensionality	Ref.
SCR	A combined deformable registration method with gray level based shortest path segmentation algorithm.	2D pixels	[3]
INR	A supervised voxelwise classification technique using layered spatio-temporal forests.	3D models	[5]
AO	A greedy optical flow algorithm with additional smoothing constraint.	2D lines/pixels	[1]
DS	A successive contour tracking algorithm based on matching correlation coefficients.	2D lines/pixels	†
EM	An active contour model framework with an optical flow energy force.	2D pixels	†
AU	An expert-guided 3D finite element heart model fitting based on guide point modeling.	3D models	-
NU	A manually expert-drawn myocardial contours.	2D contours	-

† Rater submitted the segmentation results but did not publish the corresponding methodology.

2.3 Evaluation method

Individual rater performance was measured in two aspects: (1) the accuracy of the segmentation results against the ground truth and (2) the clinical assessment of global LV mass and volume. For the rater accuracy assessment, sensitivity (p), specificity (q), positive predictive value (PPV) and negative predictive value (NPV) were the main quantitative values. These were calculated by using the following equations:

$$p = \frac{T_1}{N_1}, \quad q = \frac{T_0}{N_0}, \quad PPV = \frac{T_1}{T_1 + F_1}, \quad NPV = \frac{T_0}{T_0 + F_0} \quad (1)$$

where T_1 and T_0 are the number of detected pixels characterized correctly as myocardium and non-myocardium, while F_1 and F_0 are the number of misclassified pixels detected as myocardium and non-myocardium, respectively. The total number of myocardial and non-myocardial pixels are N_1 and N_0 , respectively.

Other commonly used evaluation metrics include similarity indices in terms of the Dice index:

$$\mathcal{D}(D_1, T_1) = \frac{2|D_1 \cap T_1|}{|D_1| + |T_1|} \quad (2)$$

and the Jaccard index:

$$\mathcal{J}(D_1, T_1) = \frac{|D_1 \cap T_1|}{|D_1 \cup T_1|} \quad (3)$$

where D_1 and T_1 are raters and ground truth sets of myocardial pixels, and $|X|$ denotes the number of elements in the set X . In both cases, values closer to 1 represent better performance.

2.4 Region of interest definition

It is well known that if $N_0 \gg N_1$, then the specificity (q) and NPV results are not particularly informative. To avoid this, we defined a region of interest

around myocardium such that N_0 is comparable with N_1 . For each image slice, a region of interest (ROI) around myocardium was defined to reduce N_0 . Two expert segmentation results (AU and NU) were added and subsequently dilated 1.5 times the width of myocardium to generate the ROI image (see Fig. 1). This produced sufficient area for each rater decision without introducing an excessive amount of background pixels. The ROI images were applied as image masks during the STAPLE iteration, as well as for the rater performance evaluation.

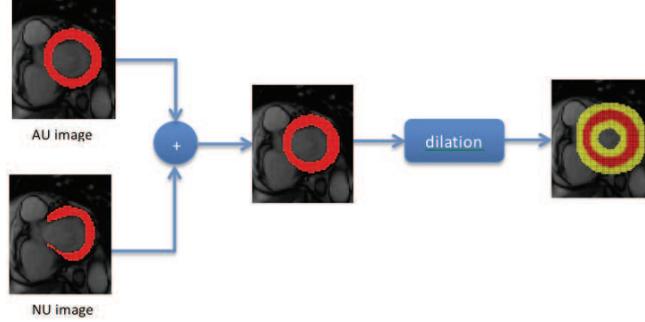


Fig. 1. A diagram to define a region of interest around the myocardium.

2.5 Binary STAPLE algorithm

In this collation study, we estimated the ground truth from all raters. Warfield et. al. [6] developed a method to estimate ground truth images from a set of segmentation results produced by raters (human and/or algorithmic) based on the EM method. The method, known as Simultaneous Truth And Performance Level Estimation or STAPLE, collects rater results and then simultaneously computes both probabilistic estimates of the true segmentation and the rater performances.

Let $\{\mathbf{D}\}_R$ be a set of R rater segmentations, each of which is an N -length of binary vector consisting of 0 (non-object) and 1 (object) values. Let \mathbf{T} be the hidden true binary vector that is going to be estimated. The objective of STAPLE algorithm is to estimate rater performance parameter $\boldsymbol{\theta}$ by maximizing the complete data log-likelihood,

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \ln f(\mathbf{D}, \mathbf{T} | \boldsymbol{\theta}). \quad (4)$$

The performance parameters are $\boldsymbol{\theta}_j = (p_j, q_j)^T$ or the sensitivity and the specificity of rater j , which can be estimated as follows

$$p_j = Pr(D_{ij} = 1 | T_i = 1) \quad (5)$$

$$q_j = Pr(D_{ij} = 0 | T_i = 0) \quad (6)$$

where $i = 1, \dots, N$ and $j = 1, \dots, R$. The parameters $p_j, q_j \in [0, 1]$ define rater performance characteristics, which are generally not equal between raters.

Applying (4) into the EM algorithm, we can define the *maximization step* as follows

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{(k)} &= \arg \max_{\boldsymbol{\theta}} E \left[\ln f(\mathbf{D}, \mathbf{T} | \boldsymbol{\theta}) | \mathbf{D}, \boldsymbol{\theta}^{(k-1)} \right] \\ &= \arg \max_{\boldsymbol{\theta}} E \left[\ln f(\mathbf{D} | \mathbf{T}, \boldsymbol{\theta}) f(\mathbf{T}) | \mathbf{D}, \boldsymbol{\theta}^{(k-1)} \right]\end{aligned}\quad (7)$$

where k denotes an iteration number and $f(\mathbf{T})$ is the stationary prior. The *expectation step* is defined by estimating the posterior probability given the current estimate of parameters at each k th iteration, i.e.,

$$f(\mathbf{T} | \mathbf{D}, \boldsymbol{\theta}^{(k)}) = \frac{f(\mathbf{D} | \mathbf{T}, \boldsymbol{\theta}^{(k)}) f(\mathbf{T})}{\sum_{\mathbf{T}} f(\mathbf{D} | \mathbf{T}, \boldsymbol{\theta}^{(k)}) f(\mathbf{T})}.\quad (8)$$

Note that the following holds for binary segmentation, i.e.,

$$f(T_i = 0 | \mathbf{D}, \boldsymbol{\theta}^{(k)}) = 1 - f(T_i = 1 | \mathbf{D}, \boldsymbol{\theta}^{(k)})\quad (9)$$

In [6], the global stationary prior $f(\mathbf{T})$ was estimated from all raters. In this study, we calculated the global stationary prior from expert raters only, i.e. AU and NU raters. Hence,

$$Pr(T_i = v) = \frac{1}{2N} \sum_{i=1}^N \sum_{j \in R': D_{ij}=v} 1\quad (10)$$

for all $i = 1, \dots, N$, $v = \{0, 1\}$ and $R' = \{\text{AU}, \text{NU}\}$.

3 Results

After the challenge submission, 18 cases had segmentations from all seven raters. Only short-axis image series at end-diastole (ED) and end-systole (ES) frames were included in the collation study, because the NU segmentations were only available at these frames. STAPLE images were estimated on each 2D image slice, independently. The total number of image slices was 330.

Clinical assessment on ED volume (EDV), ES volume (ESV) and mass were validated against the AU rater (since mass and volume were not available for the NU rater). Each automated rater provided their volume and mass estimations, while the AU volume and mass were calculated from 3D finite element models of the heart [7]. Table 3 shows the clinical assessment results in terms of mean (μ_d) and standard deviation (σ_d) of the differences.

Due to individual algorithm features, raters might not segment myocardium on a particular slice, particularly at the apical tip or basal planes. The binary

Table 3. Clinical validations on global LV functions with the AU models as the reference.

Rater	EDV diffs (ml)		ESV diffs (ml)		Mass diffs (gr)	
	μ_d	σ_d	μ_d	σ_d	μ_d	σ_d
SCR	13.03	18.13	18.98	16.20	-9.56	22.58
INR	-79.88	39.62	-61.96	38.83	74.75	53.74
AO	8.69	99.39	13.36	64.16	51.49	95.65
DS	3.94	23.14	25.65	17.71	1.08	28.40
EM	-77.75	50.60	-45.46	36.44	-51.92	39.92

Table 4. Segmentation accuracy validations with STAPLE segmentation as the reference. All numbers are in ‘average (standard deviation)’ format.

Rater	Sensitivity	PPV	Specificity	NPV	Dice	Jaccard
SCR	0.78 (0.15)	0.92 (0.07)	0.96 (0.04)	0.87 (0.08)	0.83 (0.11)	0.73 (0.14)
INR	0.75 (0.24)	0.66 (0.14)	0.73 (0.16)	0.85 (0.12)	0.68 (0.17)	0.53 (0.17)
AO	0.90 (0.12)	0.83 (0.09)	0.87 (0.09)	0.94 (0.06)	0.86 (0.09)	0.76 (0.12)
DS	0.79 (0.16)	0.82 (0.13)	0.88 (0.08)	0.87 (0.09)	0.80 (0.14)	0.68 (0.16)
EM	0.89 (0.10)	0.89 (0.09)	0.91 (0.08)	0.93 (0.06)	0.88 (0.07)	0.80 (0.10)
AU	0.85 (0.11)	0.93 (0.09)	0.96 (0.04)	0.90 (0.08)	0.88 (0.09)	0.80 (0.13)
NU	0.63 (0.12)	0.96 (0.06)	0.99 (0.02)	0.81 (0.06)	0.75 (0.10)	0.61 (0.12)

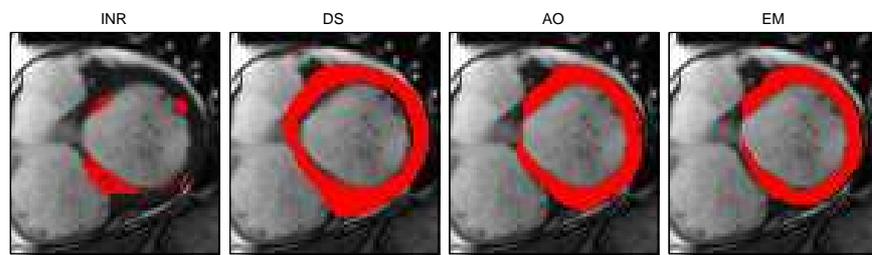
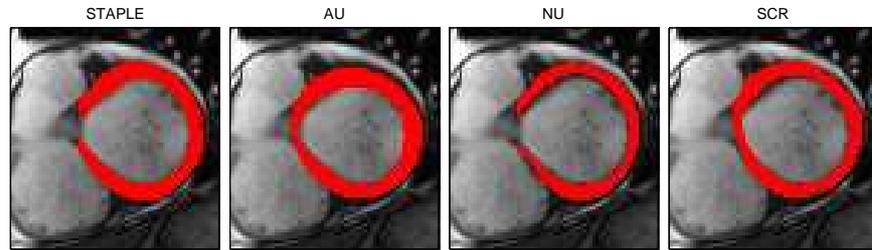
STAPLE algorithm was implemented in Matlab, based on [6]. The STAPLE algorithm was performed with the following settings: maximum of 500 iterations, a relative convergence rate of 1e-16, and the average of all raters was used as the initial weight image to define $\theta^{(0)}$. ROI images were applied.

Two examples of STAPLE images from basal and mid-ventricular slices are shown in Fig. 2. The performance of each rater is shown in Table 4. The distribution of sensitivity and specificity values are given in Fig. 3. In Fig. 4, PPV and NPV values of each rater were compared against different references: AU, NU and STAPLE. Finally, receiver operating characteristic (ROC) curves are shown in Fig. 5. Area under ROC curves (AUC) are also shown in Fig. 5.

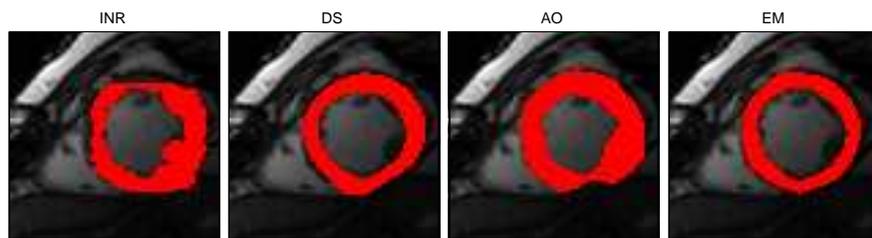
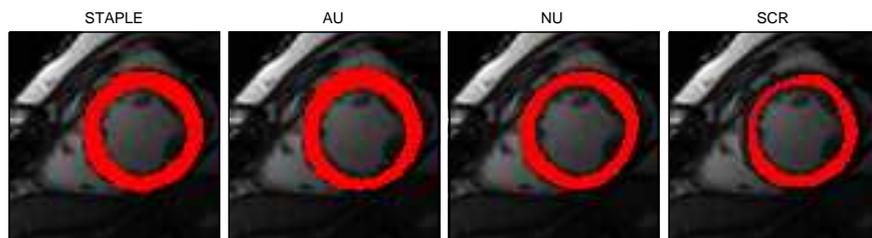
4 Discussion

A collation study from the LV segmentation challenge has been presented in this paper. A consensus segmentation was generated by using the STAPLE algorithm, which has been modified to include expert raters when estimating the global priors and to limit the segmentation area with ROI images. In general, the STAPLE algorithm produced satisfactory segmentation results, which can be regarded as the ‘ground truth’. The STAPLE method was able to resolve disagreements in the septal region of the basal plane as seen in Fig. 2(a). It also excluded papillary muscles in the mid-ventricular slices, because the majority of the raters excluded these areas from myocardium (see Fig. 2(b)).

From the clinical validation (Table 3), SCR was the closest to the expert rater AU. In terms of segmentation accuracy (Table 4), AO rater produced the highest



(a) Basal slice



(b) Mid-ventricular slice

Fig. 2. Examples of STAPLE images compared with other raters.

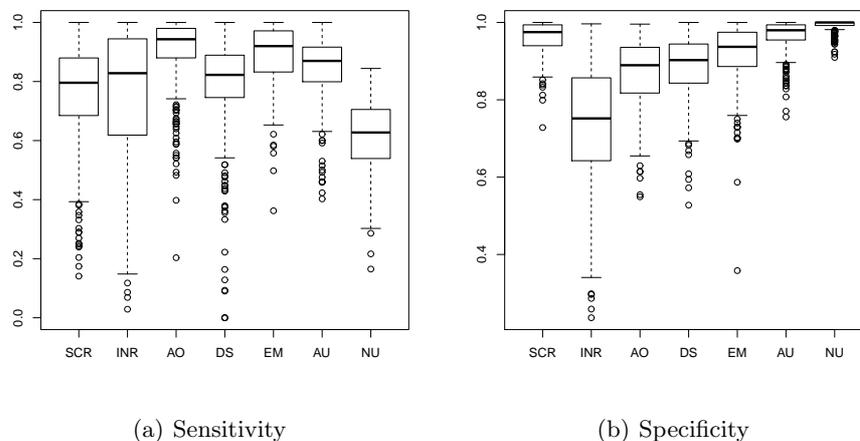


Fig. 3. Distributions of sensitivities and specificities against STAPLE images. Whiskers denote \pm interquartile-range, circles are outliers, and boxes are defined from lower to upper quartiles with median values as thick lines inside the boxes.

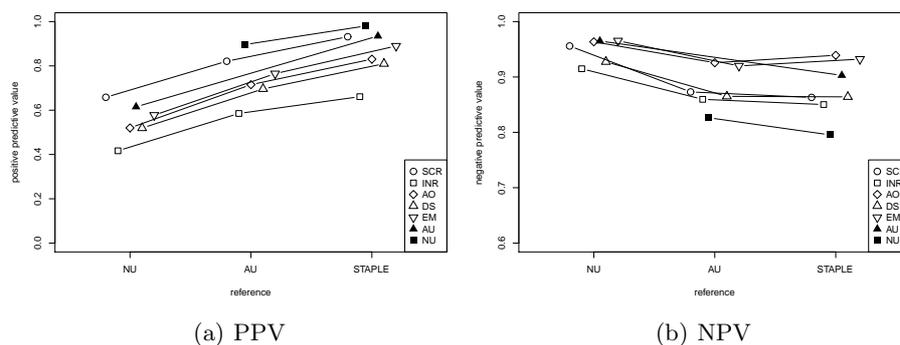


Fig. 4. Comparisons of the average PPV and NPV values by using AU, NU and STAPLE as the references.

sensitivity, while NU showed the highest predictive value. The highest similarity indices were AU and EM, both in Dice and Jaccard indices. The box plot distributions of the sensitivity and specificity values in Fig. 3 show how each rater performed. Generally, INR produced wide spread of sensitivity and specificity values, while NU maintained the highest specificity distribution values.

Figure 4 shows how PPV and NPV values varied when different expert raters were used as the reference. Applying STAPLE generally increased PPV values, but not the NPV values. The overall performances of the expert raters, as seen

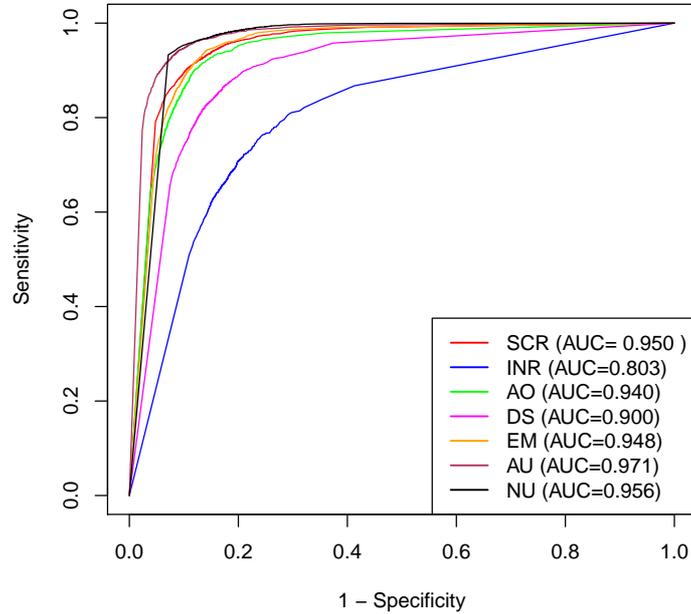


Fig. 5. Receiver operating characteristic curves from all raters. Area under the ROC curves (AUC) are captioned in the legend.

by the ROC curves in Fig. 5, were the highest among other raters. However, the expert raters did not always outperform the automated raters when compared to the STAPLE results. It is interesting to note that the expert NU rater has the lowest sensitivity among all others. The NU rater PPV values, however, were the highest. Both the lower sensitivity and higher PPV values are attributable to the smaller NU myocardium compared to other raters. This was in part due to manual exclusion of non-myocardial pixels by the NU rater, e.g., aortic root on basal slice (Fig. 2(a)) and adjacent pericardial fat on mid slice (Fig. 2(b)). The NU rater also integrated information from the entire cardiac MR study to determine the location of myocardial borders, including long-axes cine MRI and late gadolinium enhanced images. This result highlights the need for a greater consensus within the cardiac imaging community as to the acceptable criteria for accurate and reproducible segmentations.

In conclusion, STAPLE provides a mathematically objective ground truth based on the evidence from the contributing raters. This will be useful in the future to not only evaluate automated segmentation methods, but also to inform the expert decisions on what constitutes an expert consensus.

References

- [1] Fahmy, A.S., Othman, A., Khalifa, A.: Myocardial segmentation using contour-constrained optical flow tracking. In: *Statistical Atlases and Computational Models of the Heart: Imaging and Modelling Challenges* (2011)
- [2] Fonseca, C.G., Backhaus, M., Bluemke, D.A., Britten, R.D., Chung, J.D., Cowan, B.R., Dinov, I.D., Finn, J.P., Hunter, P.J., Kadish, A.H., Lee, D.C., Lima, J.A.C., Medrano-Gracia, P., Shivkumar, K., Suinesiaputra, A., Tao, W., Young, A.A.: The Cardiac Atlas Project – An imaging database for computational modeling and statistical atlases of the heart. *Bioinformatics* 27(16), 2288–2295 (2011)
- [3] Jolly, M.P., Guetter, C., Lu, X., Xue, H., Guehring, J.: Automatic segmentation of the myocardium in cine MR images using deformable registration. In: *Statistical Atlases and Computational Models of the Heart: Imaging and Modelling Challenges* (2011)
- [4] Kadish, A.H., Bello, D., Finn, J.P., Bonow, R.O., Schaechter, A., Subacius, H., Albert, C., Daubert, J.P., Fonseca, C.G., Goldberger, J.J.: Rationale and design for the Defibrillators to Reduce Risk by Magnetic Resonance Imaging Evaluation (DETERMINE) trial. *J Cardiovasc Electrophysiol* 20(9), 982–7 (2009)
- [5] Margeta, J., Geremia, E., Criminisi, A., Ayache, N.: Layered spatio-temporal forests for left ventricle segmentation from 4D cardiac MRI data. In: *Statistical Atlases and Computational Models of the Heart: Imaging and Modelling Challenges* (2011)
- [6] Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* 23(7), 903–21 (2004)
- [7] Young, A.A., Cowan, B.R., Thrupp, S.F., Hedley, W.J., Dell’Italia, L.J.: Left ventricular mass and volume: fast calculation with guide-point modeling on MR images. *Radiology* 216(2), 597–602 (2000)