Nonsupervised Ranking of Different Segmentation Approaches: Application to the Estimation of the Left Ventricular Ejection Fraction From Cardiac Cine MRI Sequences

Jessica Lebenberg, I. Buvat, Alain Lalande, Patrick Clarysse, Christopher Casta, Alexandre Cochet, Constantin Constantinidés, Jean Cousty, Alain de Cesare, Stéphanie Jehan-Besson, et al.

Nonsupervised Ranking of Different Segmentation Approaches: Application to the Estimation of the Left Ventricular Ejection Fraction From Cardiac Cine MRI Sequences

Jessica Lebenberg*, Irène Buvat, *Member, IEEE*, Alain Lalande, Patrick Clarysse, *Member, IEEE*, Christopher Casta, Alexandre Cochet, Constantin Constantinidès, Jean Cousty, Alain de Cesare, Stéphanie Jehan-Besson, Muriel Lefort, Laurent Najman, Elodie Roullot, Laurent Sarry, Christophe Tilmant, Mireille Garreau, and Frédérique Frouin, *Member, IEEE*

***Abstract*—A statistical methodology is proposed to rank several estimation methods of a relevant clinical parameter when no gold standard is available. Based on a regression without truth method, the proposed approach was applied to rank eight methods without using any *a priori* information regarding the reliability of each method and its degree of automation. It was only based on a *prior* concerning the statistical distribution of the parameter of interest in the database. The ranking of the methods relies on figures of merit derived from the regression and computed using a bootstrap process. The methodology was applied to the estimation of the left ventricular ejection fraction derived from cardiac magnetic resonance images segmented using eight approaches with different degrees of automation: three segmentations were entirely manually performed and the others were variously automated. The ranking of methods was consistent with the expected performance of the estimation methods: the most accurate estimates of the ejection fraction were obtained using manual segmentations. The robustness of the ranking was demonstrated when at least three methods were compared. These results suggest that the proposed statistical approach might be helpful to assess the performance of estimation methods on clinical data for which no gold standard is available.**

***Index Terms*—Bootstrap process, cardiac image analysis, left ventricular ejection fraction, nonsupervised segmentation methods ranking, regression without truth.**

I. Introduction

IMAGE segmentation remains a central research topic in image processing and analysis. In the medical field, it is often a necessary step for other processing tasks, such as image registration, volumetric and functional analysis, or derivation of clinical parameters useful for the diagnosis. The most frequently used method consists in manually outlining regions of interest on images. However, this approach is labor intensive, time consuming and subject to intra- and inter-operator variabilities. To overcome these limitations, segmentation algorithms are developed with different degrees of automation. An evaluation of the accuracy of these automatic segmentation methods must be performed before any clinical application. In most cases, the accuracy of the resulting segmentation is assessed by comparing the estimated contours with a reference contour ideally provided by one (or several) expert(s). Evaluation can be achieved by visually inspecting the superimposition of the contours. This approach being operator-dependent, a quantitative assessment thus appears preferable. Distances between contours can be computed [1], [2]. A criterion measuring the overlap between the segmented region and the gold standard region can also be calculated, such as the Dice coefficient [3]. Such global coefficients are useful to evaluate the accuracy of segmentation results with respect to a reference segmentation. However, the gold standard is not always easily available. Furthermore, since small errors in delineations may have an important impact on further image processing steps, possibly yielding misleading results and/or leading to inappropriate treatment [4], [5], it is essential to propose new solutions for comparing the performance of segmentation methods when the ground truth is unknown. In addition, it appears relevant to

Manuscript received February 29, 2012; revised May 16, 2012; accepted May 16, 2012. Date of publication May 30, 2012; date of current version July 27, 2012. This work was performed in the framework of the French MedIEval (**Med**ical **I**mage segmentation **Eval**uation) working group. This work was supported by GdR 2647 Stic-Santé under its support of the MedIEval action. *Asterisk indicates corresponding author.*

*J. Lebenberg is with the LIF, INSERM UMR_S 678, Université Pierre et Marie Curie, 75013 Paris, France, and also with the PRIAM, ESME-Sudria, 94200 Ivry-sur-Seine, France (e-mail: jessica.lebenberg@gmail.com).

C. Constantinidès is with the LIF, INSERM UMR_S 678, Université Pierre et Marie Curie, 75013 Paris, France, and also with the PRIAM, ESME-Sudria, 94200 Ivry-sur-Seine, France.

A. de Cesare, M. Lefort, and F. Frouin are with the LIF, INSERM UMR_S 678, Université Pierre et Marie Curie, 75013 Paris, France (e-mail: frouin@imed.jussieu.fr).

E. Roullot is with the PRIAM, ESME-Sudria, 94200 Ivry-sur-Seine, France.

I. Buvat is with the IMNC, CNRS UMR 8165, Université Paris-Sud, 91405 Orsay, France.

A. Lalande and A. Cochet are with the Le2I, CNRS UMR 6306, Université de Bourgogne,71200 Dijon, France.

P. Clarysse and C. Casta are with the Université de Lyon, Creatis, CNRS UMR5220, Inserm U1044, INSA-Lyon, Université de Lyon 1, 69621 Villeurbanne, France.

J. Cousty and L. Najman are with the LIGM-A3SI, UMR 8049, Université Paris-Est, ESIEE, 77454 Marne la Vallée, France.

S. Jehan-Besson is with GREYC, CNRS UMR 6072, 14032 Caen, France.

L. Sarry is with ISIT, CNRS UMR 6284, Université d'Auvergne, 63000 Clermont-Ferrand, France.

C. Tilmant is with the Institut Pascal, CNRS UMR 6602, Université Blaise Pascal, 63000 Clermont-Ferrand, France .

M. Garreau is with the LTSI, INSERM UMR 1099, Université Rennes 1, 35042 Rennes, France.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2012.2201737

evaluate the segmentation approaches with respect to the real user objective.

A method to compare two measurements was proposed by Bland and Altman [6]. Their approach aims at assessing the agreement between two methods. It consists in plotting the mean of the two parameter estimates versus the difference between both parameter estimates. As the mean of differences can represent the mean bias between the two methods and the standard deviation is related to the data variability, it is possible to test whether the two methods under evaluation provide similar results.

When several segmentation methods have to be evaluated, the aforementioned approach quickly becomes tedious and does not allow the ranking of the methods. To address this issue, the "Regression Without Truth" approach (RWT) was proposed [4], [7], [8]. The proposed method consisted in estimating the parameters of a linear model establishing the relationship between the measurements and the clinical index of interest. Among the hypotheses used for the ranking of the different segmentation approaches, it was assumed that the distribution of the true value of the clinical parameter of interest follows a finite support distribution. The parameters of this distribution were estimated jointly with regression parameters linking the estimated clinical indices to the true values. A figure of merit characterizing the performance of each estimation method was then derived. In their first reports, the authors studied two finite support distributions (a Beta and a truncated normal distributions) to describe 100 ejection fractions (EF) simulated for the study.

On the basis of the previous work, Kupinski *et al.* compared the results provided by three different algorithms for EF estimation [4]. This study was performed on single-photon emission computed Tomographic images acquired on 100 patients. In this study, a comparison between the Beta and the truncated normal distributions was performed leading to the choice of the Beta distribution which seemed to be the most appropriate for studying clinical parameters with bounded values such as the EF. Jha *et al.* recently used the RWT approach to compare the performance of three segmentation algorithms developed to estimate the apparent diffusion coefficient of lesions from diffusion-weighted magnetic resonance images (MRI) [9], assuming that these coefficients followed a Beta distribution. The parameters of this finite support distribution were estimated among a reasonable range of values during the comparison process. The authors conclusions suggested to consider *prior* information regarding the finite support distribution to improve the reliability of the ranking of the segmentation approaches.

In this paper, we propose a methodology based on the RWT approach to rank segmentation methods with different degrees of automation. This extended RWT (eRWT) approach was carried out using a figure of merit introduced in [10] which differed from the one initially introduced in [7] and [8]. To get a robust comparison, a bootstrap analysis [5], [11] was performed on top of the eRWT approach, followed by a rank analysis. A preliminary study was performed using *prior* information regarding the finite support distribution describing the statistical distribution of the EF of the left ventricle (LV) in a database [12]. This work was completed here in assessing the robustness and limitations of the ranking. The influence of several parameters was studied, especially the parameters of the finite support distribution describing the studied clinical parameter of interest. We also investigated the impact of the number of segmentation approaches to be compared. The method is here illustrated in the particular context of the EF estimation in cardiac cine MRI sequences with up to eight endocardial segmentation approaches available from different research teams.

This paper is organized as follows. Section II presents the database used in our study and the segmentation methods to be compared. Section III explains the eRWT theory. Section IV describes the experiments performed to assess the robustness and limitations of eRWT. Section V shows the ranking of the segmentation methods and the results regarding the robustness of the eRWT method. The method and results are discussed in Section VI.

II. Database and Segmentation Methods

A. Database

The eRWT was applied to the EF estimated from the segmentation of the MRI cardiac datasets provided to the participants in the MICCAI 2009 Grand Challenge by Sunnybrook Health Sciences Center [13]. The database consisted of the stacks of sequences corresponding to 45 subjects from the training, the testing and the online contest datasets. It included nine healthy individuals and 36 patients with various cardiac pathologies: 12 hypertrophic cardiomyopathy, 12 heart failure without ischemia, and 12 heart failure due to ischemia. For each patient, about ten MR short axis slices covering the LV were acquired using a Steady State Free Precession sequence. Twenty phases covering the cardiac cycle were acquired and the acquisition was triggered by the R wave of the ECG (the first phase corresponds to the end-diastole). Further details regarding datasets and image acquisition protocol can be found in [13].

The contestants were challenged to provide the best possible segmentation and the most reliable EF estimate with respect to a manual reference. The EF biomarker is defined as the ratio between the end-diastolic and the end-systolic LV volumes difference and the end-diastolic volume. It ranges from 0 to 1. A score higher than 0.5 is considered as normal. The 24 patients of the studied database with heart failure had a reduced EF (≤ 0.45). More than 99% of EFs ranged from 0.05 to 0.85.

MR slices corresponding to the end-systolic and end-diastolic phases were indicated to the participants in the challenge, so as to avoid any variability due to the choice of these temporal phases.

B. Segmentation Approaches

Eight segmentation approaches proposed by five different research teams were included resulting in eight independent estimates of the EF.

Methods M1, M2, and M3 were entirely manual and were provided by three experts from two laboratories. Semi-automated methods M4, M5, and M6, described in [14]–[16], respectively, involved an interactive definition of an initial shape or a modification of the parameters used by the operators

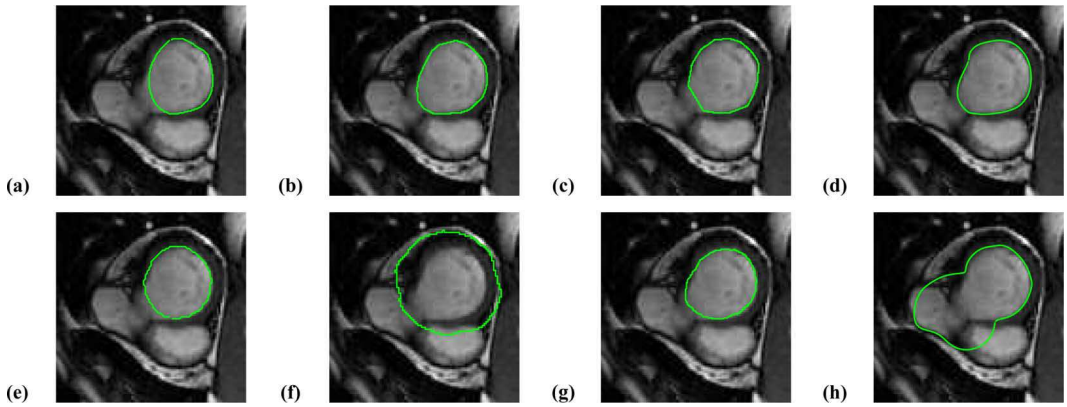

Fig. 1. Basal cine MRI slice from a patient at end-diastole with superimposed contours of the LV (green line) provided by the eight segmentation methods included in our study (M1 to M8 represented from (a) to (h), respectively).

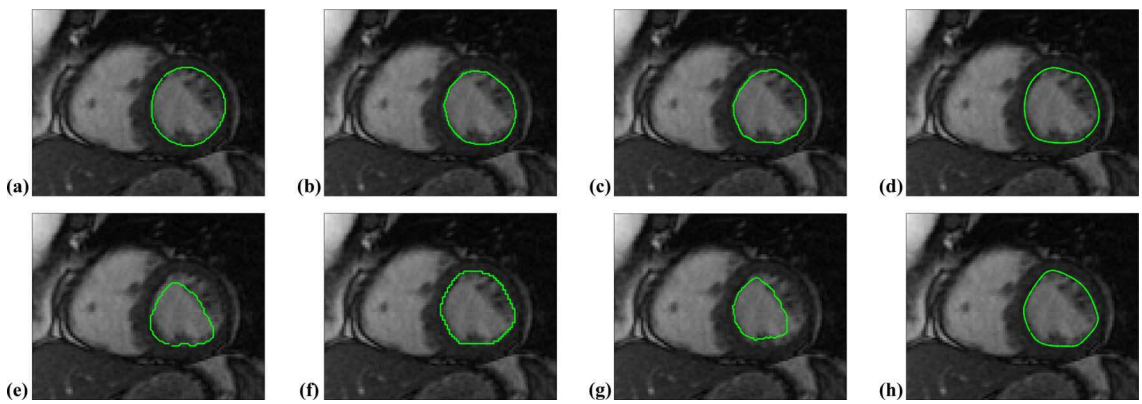

Fig. 2. Median cine MRI slice from a healthy individual at end-diastole with superimposed contours of the LV (green line) provided by the eight segmentation methods included in our study (M1 to M8 represented from (a) to (h), respectively).

during the segmentation process. Method M7, described in [17], was mostly automated. Method M8 was fully automated and its main principle was described in [18]. All of the aforementioned methods included the papillary muscles (PM) in the LV cavity except for M5. All of them were applied to the end-systolic and end-diastolic phases (3D segmentation) except for M5 and M6 that provided LV contours for the whole temporal sequence (4D segmentation). For the latter, only the end-systolic and end-diastolic results were taken into account in this study.

Figs. 1 and 2 give examples of endocardial contours, obtained by each segmentation approach, superimposed on a basal and on a median MR end-diastolic image from a patient and a healthy individual of the database.

III. Extended Regression Without Truth (eRWT) Approach

A. Theory

The RWT approach, detailed in [4], [7], [8], is only summarized here.

Let us consider the database containing images from P patients (indexed by p, ranging from 1 to P) and M segmentation methods (indexed by m, ranging from 1 to M). Each segmentation method m yields an estimate θ_{pm} of the biomarker of interest on the sample p. The true value Θ_p of this biomarker is unknown.

The RWT approach assumes a parametric relationship between the true value Θ_p and its estimate θ_{pm} based on the three following hypotheses.

H1: The statistical distribution of the true value Θ_p for the whole database has a finite support.

H2: Each method m provides an estimate θ_{pm} of Θ_p which is linearly related to Θ_p (1) where the error term ε_{pm} is normally distributed with zero mean and standard deviation σ_m, and where the a_m and b_m parameters are specific to each method m and independent of the sample p

$$\theta_{pm} = a_m \Theta_p + b_m + \varepsilon_{pm}. \tag{1}$$

H3: The error terms ε_{pm} for each method m are independent.

Given these assumptions, the probability of the estimated values θ_{pm} given the linear model and the true value Θ_p can be expressed by (2)

$$pr(\{\theta_{pm}\}|\{a_m, b_m, \sigma_m^2\}, \Theta_p) = \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left(-\frac{1}{2\sigma_m^2}\left(\theta_{pm} - a_m\Theta_p - b_m\right)^2\right). \tag{2}$$

Considering the P samples of the database, the log-likelihood $\ln(L)$ can be written, using (3), as a function of a_m, b_m, and σ_m and the probability of Θ_p

$$\ln(L) = P \ln\left(\prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma_m^2}}\right) + \sum_{p=1}^{P} \ln\left[\int \mathrm{d}\Theta_p pr\left(\Theta_p\right) \times \exp\left(\sum_{m=1}^{M}\left(-\frac{1}{2\sigma_m^2}\left(\theta_{pm} - a_m\Theta_p - b_m\right)^2\right)\right)\right]. \quad (3)$$

The maximization of this likelihood does not require the numerical values of Θ_p, but only a model of its statistical distribution $(pr\left(\Theta_p\right))$ and leads to the estimates of the linear model parameters for each method (a_m, b_m, and σ_m).

According to [4], the finite support Beta distribution is well-appropriated to describe the distribution of the biomarkers Θ_p (such as the EF) ranging between two fixed values. This distribution was therefore chosen in our study. The log-likelihood defined in (3) can then be computed using the probability density function of a Beta distribution given by the following equation (4):

$$pr(\Theta_p) = \frac{\Theta_p^{\mu-1}(1-\Theta_p)^{\nu-1}}{B(\mu,\nu)}, \quad \text{with } B(\mu,\nu) = \int_0^1 x^{\mu-1}(1-x)^{\nu-1}\,\mathrm{d}x. \quad (4)$$

In this work, the estimation of the linear model parameters was performed by numerically optimizing a constrained nonlinear multivariable function implemented in MATLAB (R2009a, The Mathworks, Inc.) and based on a sequential quadratic programming method [19]. To ensure convergence of the optimization algorithm, the constrained nonlinear multivariable function was initialized close to the expected solution, i.e., with a slope close to 1, an intercept and a standard deviation of the error term close to 0.

B. Figure of Merit

The figure of merit proposed in [4], [7] for comparing the different methods was the ratio between σ_m and a_m: the smaller this ratio, the more accurate the estimation method. According to this criterion, if two segmentation approaches estimate the clinical parameter of interest with a similar standard deviation of the error term, the most accurate method will be the one with the highest a_m slope.

Equation (1), however, suggests that Θ_p is accurately estimated if the slope a_m is close to 1, the intercept b_m close to 0, and the error term has a low standard deviation. To quantify the deviation between the estimated values and the "ideal" estimates represented by the identity line, Soret *et al.* introduced in [10] another figure of merit F_m defined as the mean square difference between the true value of the parameter and the estimated value [(5)]

$$F_m = \mathbb{E}\left[\left(\Theta - a_m\Theta - b_m - \varepsilon_m\right)^2\right]. \quad (5)$$

As the mean $(\mathbb{E}\left[\Theta\right])$ and the variance $(\mathrm{Var}\left[\Theta\right])$ of a Beta distribution are given by (6) and (7), respectively, the second moment $(\mathbb{E}\left[\Theta^2\right])$ is given by (8)

$$\mathbb{E}\left[\Theta\right] = \frac{\mu}{\mu+\nu} \quad (6)$$

$$\mathrm{Var}\left[\Theta\right] = \frac{\mu\nu}{(\mu+\nu)^2(\mu+\nu+1)} \quad (7)$$

$$\text{and } \mathbb{E}\left[\Theta^2\right] = \mathrm{Var}\left[\Theta\right] + \mathbb{E}\left[\Theta\right]^2 = \frac{\mu}{\mu+\nu} \times \frac{\mu+1}{\mu+\nu+1}. \quad (8)$$

Using hypotheses **H1**, **H2**, and **H3**, F_m can thus be expressed using (9)

$$F_m = \left(a_m - 1\right)^2 \frac{\mu\left(\mu+1\right)}{\left(\mu+\nu\right)\left(\mu+\nu+1\right)} + 2\left(a_m - 1\right) b_m \frac{\mu}{\mu+\nu} + b_m^2 + \sigma_m^2. \quad (9)$$

The ranking of the segmentation methods using the eRWT is based on this figure of merit.

C. Bootstrap Process and Rank Analysis in the eRWT Approach

To get robust estimates of F_m and overcome robustness issues due to low sample size, a bootstrap approach was used. This statistical process is extensively described in [11]. The principle consists in randomly drawing with replacement P samples from the initial sample, and repeat the process N times.

For each drawing n, the parameters a_m^n, b_m^n and σ_m^n were estimated, yielding a figure of merit F_m^n. The nonparametric Kruskal-Wallis test [20] was then performed based on the $N \cdot M$ figures of merit F_m^n to determine whether the median figure of merit was equal among segmentation methods. When the null hypothesis was rejected, the mean rank of each method was compared two by two, using a Bonferroni correction with a Type I error equal to 5% [21].

IV. Experiments

A. Comparison of the Segmentation Approaches Using the eRWT Methodology

1) Estimation of the eRWT Parameters: A first study was performed by setting the parameters of the Beta distribution to fulfill two criteria:

C1: since there were more subjects with a reduced EF than with a normal EF, the distribution was centered at a value less than 0.5;

C2: since more than 99% of the EFs ranged from 0.05 to 0.85, μ and ν were chosen so that the probability density function was close to 0 outside this range.

The μ and ν parameters of the Beta distribution representative of the database were empirically determined. Parameters μ and ν, respectively, equal to 4 and 5 met observations **C1** and **C2**. The probability density function of such a Beta distribution is represented in Fig. 3(a).

The linear model parameters were first estimated without considering a bootstrap. The figure of merit F_m was computed

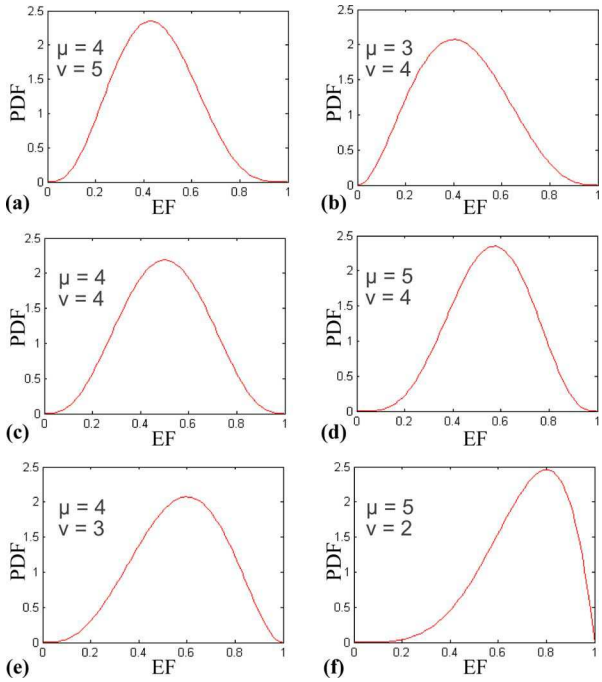

Fig. 3. Probability density functions of Beta distributions corresponding to different μ and ν parameters.

for each method and a preliminary ranking of the segmentation methods was deduced.

2) Analysis of F_m After Bootstrap and Rank Analysis: To get robust results, the bootstrap approach was applied. In the present work, $N = 1000$ different random drawings were performed from the $P = 45$ initial samples. As $M = 8$ segmentation methods were compared, the rank analysis was performed based on 8000 figures of merit.

The distribution of F_m for each segmentation method was studied.

B. Robustness of eRWT

The bootstrap process considering $N = 1000$ different random drawings was systematically applied to the following experiments.

1) Influence of the Finite Support Distribution Parameters: In the first studies regarding the RWT approach applied to the EF estimates [7], the parameters of the finite support distribution were constrained to lie between 1 and 5 during the RWT optimization process. Yet, Jha *et al.* suggested in the conclusion of their work described in [9] that estimating the Beta distribution parameters during the RWT process was not necessarily the best approach. To further investigate the role of the parameters of the finite support distribution, we compared the segmentation approaches as described in previous papers, i.e., in optimizing the log-likelihood given by (3) so that the estimated Beta distribution parameters were in the $[1, 5]$ range.

The eRWT approach was thereafter applied while setting the parameters of the Beta distribution differently. Tests were performed using other parameter pairs representing a distribution of values close to that shown in Fig. 3(a) [$(\mu = 3; \nu = 4)$ represented in Fig. 3(b)], a symmetric distribution [$(\mu = 4; \nu = 4)$ represented in Fig. 3(c)], and distributions with more subjects with a normal EF than with a reduced one [$(\mu = 5; \nu = 4)$, $(\mu = 4;\ \nu = 3)$ and $(\mu = 5;\ \nu = 2)$ represented in Fig. 3(d)–(f), respectively].

For each combination of parameters, F_m was estimated. The standard deviations σ_m of the error term ε_{pm} were also compared through the tests to study the σ_m robustness. A rank analysis was then performed using the values of F_m.

2) Influence of the Segmentation Approaches Included in the Comparison: To assess the robustness of the proposed comparison methodology with respect to the segmentation approaches under evaluation, the ranking of the methods was performed considering different combinations of segmentation approaches:

- semi- and largely automated approaches (i.e., methods M4, M5, M6, M7, and M8);
- manual and largely automated approaches (i.e., methods M1, M2, M3, M7, and M8);
- manual and semi-automated approaches (i.e., methods M1, M2, M3, M4, M5, and M6);
- all combinations of three methods among the eight available (number of considered combinations: $\binom{8}{3} = 56$);
- all combinations of two methods among the eight available (number of considered combinations: $\binom{8}{2} = 28$).

To assess the modification of the ranking of methods with respect to the methods entering the comparison, a ranking inversion cost $c_{i,j}$ was computed. This cost is described in (10) with r_i and r_j the reference rankings of M_i and M_j, respectively, and s_i and s_j the rankings of M_i and M_j in the current comparison study

$$\begin{cases} c_{i,j} = 0, & \text{if} \begin{cases} r_i < r_j \text{ and } s_i < s_j \\ \text{or} \\ r_i > r_j \text{ and } s_i > s_j \end{cases} \\ c_{i,j} = \mid r_i - r_j \mid + \mid s_i - s_j \mid, & \text{otherwise.} \end{cases} \tag{10}$$

This ranking inversion cost was computed for all combinations of two methods M_i and M_j among the k methods entering the comparison (number of considered combinations: $\binom{k}{2} = k!/2!(k-2)!$). This cost penalizes more a ranking inversion between two methods performing very differently in the reference ranking than a ranking inversion between two methods ranked consecutively in the reference ranking.

For this robustness study, the ranking obtained with the eight methods entering the comparison was considered as reference and the Beta distribution parameters μ and ν were set to 4 and 5, respectively.

3) Influence of the Database Size: To evaluate the proposed comparison methodology robustness with respect to the size of the database, the ranking of the methods was performed using only the testing and the online contest datasets. This smaller database consisted of 30 subjects including six healthy subjects, eight patients with hypertrophic cardiomyopathy, eight with heart failure without ischemia, and eight with heart failure due to ischemia. The sixteen patients with heart failure had a reduced EF.

The eRWT was then applied to this reduced database by including the eight methods and all combinations of three

TABLE I
ESTIMATED eRWT PARAMETERS (a_m, b_m, σ_m AND F_m) FOR EACH METHOD USING $\mu = 4$ AND $\nu = 5$ AS BETA DISTRIBUTION PARAMETERS

Method	a_m	b_m	σ_m	F_m
M1	1.198	-0.113	0.042	0.003
M2	1.245	-0.103	0.012	0.002
M3	1.305	-0.111	0.022	0.003
M4	1.270	-0.114	0.061	0.006
M5	0.914	-0.031	0.079	0.011
M6	1.431	-0.141	0.066	0.011
M7	1.148	-0.093	0.086	0.009
M8	1.145	-0.051	0.134	0.019

methods. The ranking inversion cost evaluating the modification of the ranking of methods when going from eight to three methods was also estimated using this reduced database and considering the ranking obtained with the eight methods included as the reference [cf. (10)].

To be consistent with the previous robustness study, the Beta distribution parameters μ and ν were set to 4 and 5, respectively.

V. RESULTS

A. Comparison of the Segmentation Approaches Using the eRWT Methodology

1) Estimation of the eRWT Parameters: Table I lists the parameters of the linear model (a_m, b_m and σ_m) estimated for each method using the eRWT approach with the μ and ν Beta distribution parameters set to 4 and 5, respectively, and without any bootstrap procedure.

For each method, estimates of EF with associated standard deviation were plotted against the true value of the EF ranging from 0 to 1 (see Fig. 4). A line corresponding to the "ideal" estimates (identity between the estimated values and the true value) was superimposed to these graphs. None of the methods yields estimated values very far from the identity line. The smaller the deviation between the experimental and the ideal lines along the range of possible values and the smaller the standard deviation, the more accurate the estimate. Table I and Fig. 4 suggest that the estimates of the biomarker provided by manual methods (M1, M2, and M3) were the most accurate and least variable, whereas results obtained using method M8 were more scattered. Method M5 globally underestimated EF.

The fifth column of Table I gives F_m, the figures of merit for each method. Based on the values, the ranking of the methods, beginning from the most accurate, is: M2, M1-M3, M4, M7, M5-M6, and M8.

2) Analysis of F_m After Bootstrap Process and Rank Analysis: Fig. 5 illustrates the distribution of F_m computed for each segmentation approach after the bootstrap process. This figure shows that results obtained with M8 are the most scattered and those obtained using M1, M2, and M3 are the least dispersed.

Results of the rank analysis based on F_m distinguished six groups of methods classified in descending order of performance: M2, M1-M3, M4, M7, M5-M6, and M8.

The time computation used to perform this test on a Dell Precision PWS380 computer (Windows XP, Pentium 4, 3 GHz, 3 Gb RAM) was about 3 h 30 min. This was not optimized as all routines were coded in Matlab.

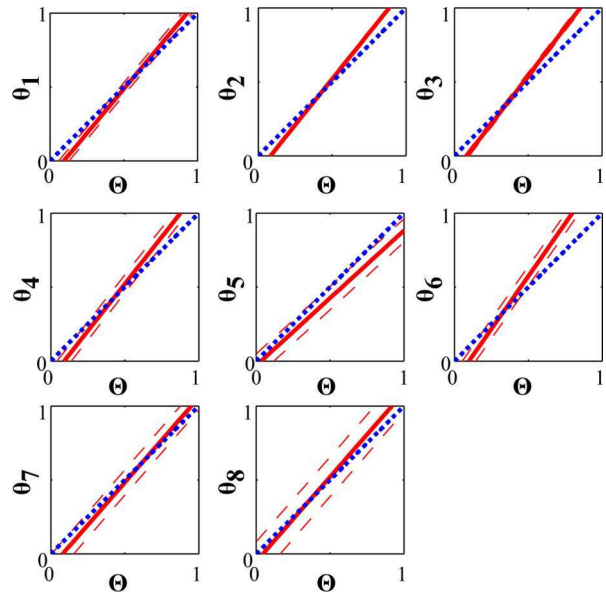

Fig. 4. Plot of the estimated EF (θ_m) as a function of the true EF (Θ) for each method (solid red line) with associated standard deviation (dashed red lines), and ideal estimates (dotted blue line= identity between evaluated values and the true value of EF).

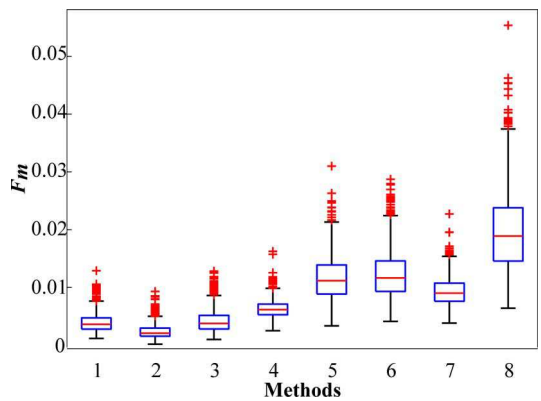

Fig. 5. Boxplot of the distribution of F_m computed after the bootstrap process for each method: the median value is represented by the red horizontal segment, the interquartile range by the blue rectangle, adjacent values inferior to 1.5 times the interquartile range by the dashed black line and outliers by red crosses. The first three methods provided the three lowest values of F_m and are the least dispersed.

B. Robustness of eRWT

1) Influence of the Finite Support Distribution Parameters: Fig. 6(a) shows, for each segmentation method, the median value of the figures of merit computed during the bootstrap process with associated first and third quartiles for different parameters of the Beta distribution. Results obtained for ($\mu = 3; \nu = 4$)—solid red line—and for ($\mu = 4; \nu = 5$)—dashed green line—are very similar, and so are the results obtained with ($\mu = 4; \nu = 3$)—dotted light blue line—and ($\mu = 5; \nu = 4$)—pink solid line with circle markers. This figure also shows that, for all segmentation methods, F_m was the lowest when μ and ν were set so that the maximum of the finite support distribution was around 0.45, i.e., for ($\mu = 3; \nu = 4$) and for ($\mu = 4; \nu = 5$). The largest F_m were obtained for parameters representing a distribution with

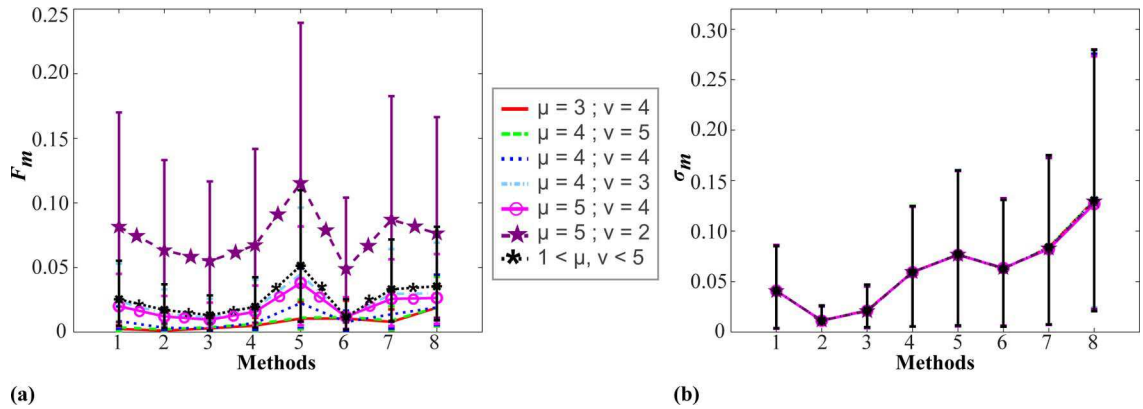

Fig. 6. (a) Median values of F_m and (b) of σ_m in the eRWT linear model computed during the bootstrap process and associated first and third quartiles for different parameters of the Beta distribution. The smallest F_m values were obtained for $(\mu = 3; \nu = 4)$ and for $(\mu = 4; \nu = 5)$, the largest F_m values for $(\mu = 5; \nu = 2)$. The term σ_m did not depend on the Beta distribution parameters (similar estimates whatever the Beta distribution parameters).

TABLE II
RANKING OF THE SEGMENTATION METHODS FOR DIFFERENT PARAMETERS OF THE BETA DISTRIBUTION. THE RANKING STRONGLY DEPENDS ON THE SET OF $(\mu;\nu)$ PARAMETERS

	$(\mu;\nu)$	(3;4)	(4;5)	(4;4)	(4;3)	(5;4)	(5;2)	[1,5]
- Performance +		M2	M2	M2-M3	M3-M6	M3	M6	M6
		M1	M1-M3			M2-M6	M3	M3
		M3		M1-M4-M6	M2		M2	M2
		M4	M4		M4	M4	M4	M4
		M7	M7		M1	M1	M8	M1
		M5-M6	M5-M6	M7	M7-M8	M7-M8	M1	M7
				M8			M7	M8
		M8	M8	M5	M5	M5	M5	M5

TABLE III
RANKING OF THE SEGMENTATION METHODS WHEN DIFFERENT COMBINATIONS OF METHODS ENTERED THE COMPARISON

		Methods entering the comparison			
	Rank number	All	Manual and semi-auto	Manual and largely auto	Semi- and largely auto
- Performance +	1	M2	M2	M2	M4
	2	M1-M3	M1-M3	M1-M3	M7
	3				**M5**
	4	M4	M4	M7	**M6**
	5	M7	M5-M6	M8	M8
	6	M5-M6			
	7				
	8	M8			

a maximum close to 0.80, i.e., for $(\mu = 5; \nu = 2)$—dashed purple line pentagram markers. This configuration apart, the highest figures of merit were obtained when the Beta distribution parameters were estimated during the eRWT process and under the constraint to lie between 1 and 5 (dotted black line with asterisk markers). This experiment shows that the joint estimation of the Beta distribution parameters and of the linear model parameters does not provide the smallest values of F_m.

Fig. 6(b) presents, for the eight segmentation methods and for the different sets of parameters of the Beta distribution, the median and associated first and third quartiles values of the standard deviations σ_m obtained during the bootstrap process. For a given method, the values of σ_m were similar whatever the Beta distribution parameters.

Table II presents the ranking of the segmentation methods for the different sets of $(\mu;\nu)$ parameters of the Beta distribution. When μ and ν yielded close probability density functions ($(\mu = 3; \nu = 4)/(\mu = 4; \nu = 5)$ and $(\mu = 4; \nu = 3)/(\mu = 5; \nu = 4)$), the ranking of methods was quite close. However, this table shows that the ranking strongly depends on the set of $(\mu;\nu)$ parameters.

2) Influence of the Segmentation Approaches Included in the Comparison: Table III presents the ranking of the segmentation methods when different combinations of methods were included in the comparison. This table shows that, when only manual and semi-automated or manual and largely automated approaches were included in the comparison, the segmentation methods ranking was identical to the reference ranking (the one obtained when all segmentation methods were included in the comparison). Thus, in both cases, 100% of the ranking inversion costs computed were equal to 0. When only semi- and largely automated methods were included in the comparison, there was no ranking inversion. Nevertheless, method M5 was ranked third and method M6 ranked fourth. As both methods had the same ranking in the reference ranking (rank number of 6.5), the ranking inversion cost computed for this ranking modification was equal to 1. All other computed ranking inversion costs were equal to 0.

Fig. 7 presents the frequency of computed ranking inversion costs to assess the modification of the ranking when only three or two methods are compared instead of eight methods (reference ranking). For each case, the frequency was normalized by the number of combinations available (case with three methods including in the comparison: $\binom{8}{3} \cdot \binom{3}{2} = 168$, case with two methods including in the comparison: $\binom{8}{2} \cdot \binom{2}{2} = 28$). When considering all combinations of three segmentation methods, 93% of computed ranking inversion costs were equal to 0 (black bar on Fig. 7). Five percent of the computed ranking inversion costs were equal to 1. The remaining cases were equal to 1.5, 2, or 2.5.

When considering all combinations of two segmentation methods, 46% of computed ranking inversion costs were equal to 0 (red bar on Fig. 7). About 11% of computed ranking inversion costs were equal to 1. The remaining 43% were either higher than or equal to a score of 2.5. This means that almost

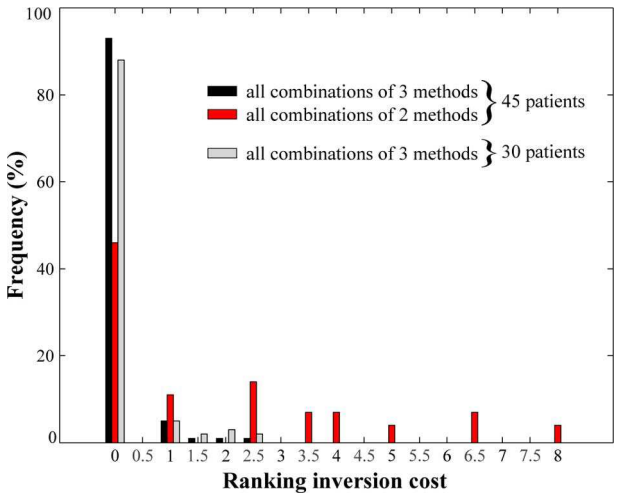

Fig. 7. Ranking inversion costs measuring the change in ranking of methods as a function of the methods included in the comparison. The reference ranking is the one obtained when including all eight methods in the comparison with the 45 patients database (red and black bars) and with the 30 patients database (gray bar).

half of the comparisons including only two segmentation methods changed the ranking of methods not consecutively ranked in the reference ranking.

3) Influence of the Database Size: Results of the rank analysis considering the reduced database (30 patients) and the eight available segmentation methods distinguished six groups of methods classified in descending order of performance: M2, M1, M3, M4-M7, M5-M6, and M8.

Eighty-eight percent of the $\binom{8}{3} \cdot \binom{3}{2}$ comparisons performed using only three methods identically ranked the segmentation methods as in the comparison involving all methods (ranking inversion cost = 0, light gray bar on Fig. 7). Five percent of the computed ranking inversion costs were equal to 1. The remaining 7% were equal to 1.5, 2, or 2.5.

VI. Discussion

The main goal of this study was to propose and characterize the performance of a comparison methodology which aims at ranking, for a database, several estimation methods when the gold standard is unknown. In our application, we ranked eight segmentation approaches with various degrees of automation used to assess the EF for nine healthy subjects and 36 patients with various cardiac pathologies. One of the segmentation methods did not include PM in the LV cavity to delineate contours. It should be emphasized that the comparison did not actually assess the segmentation accuracy, but the ability of the segmentation algorithm to properly estimate the EF, which is different as a systematic segmentation error can still lead to reasonably accurate EF.

An approach often used to assess segmentation methods consists in comparing the methods one by one by referring to a gold standard commonly provided by a manual delineation. Since manual delineation is an intensive work impractical on large databases and since the true value of the clinical parameter of interest remains unknown, a comparison approach without using any gold standard was proposed [7]. This RWT approach is based on several assumptions regarding the estimation methods under assessment. The ranking of methods is based on the computation of a figure of merit that is a function of the estimated regression parameters.

In this work, we extended this method in two respects. First we did not use the proposed figure of merit, but rather the one introduced by Soret *et al.* [10]. Second, the ranking was defined by a bootstrap approach and associated rank analysis, which allowed us to better characterize the stability of the ranking. This extended approach, named eRWT, was applied to the database provided to the participants in the MICCAI 2009 Grand Challenge to characterize its performance and evaluate its robustness with respect to various parameters. In particular, we studied how the final ranking provided by the eRWT approach depended on:

1) distribution parameters used to describe the statistical distribution of the clinical parameter of interest;
2) number of methods included in the comparison;
3) size of the database.

A. Comparison of Segmentation Methods Using the eRWT Approach

To assess the consistency of the ranking provided by the eRWT, we included three manual segmentation methods among the eight methods to be evaluated. At that point, it is important to underline that if manual delineation is often considered to be the "gold standard" when assessing segmentation methods, in our study, the manual segmentation methods did not have any *a priori* specific role. Yet, results presented in Section V-A showed that the manual delineations appeared to provide the most accurate EF estimations. This is consistent with our expectation that these methods would outperform semi-automated or largely automated analyses. Thus, the eRWT approach, only based on hypotheses described in Section III-A and using no *priors* regarding the features of the methods (manual, semi-automated, largely automated) seems to be relevant to objectively compare several segmentation approaches for subsequently deriving the EF.

Estimated eRWT parameters for each method showed that semi and largely-automated methods still provided reasonably accurate EF estimates: the slope and intercept of the linear model were close to 1 and 0, respectively, and the figures of merit were close to the scores obtained for the manual delineations. These observations were also true for M5, a method that did not include the PM in the LV cavity. This means that including or excluding the PM from the LV cavity does not seem to have a large influence on the accuracy of EF estimates. Method M8 was found to have more scattered performances than the others. This is due to the image segmentation approach used to estimate contours [18]. Indeed, a region of interest around the LV was automatically computed using a circular Hough transform. Then, a morphological filtering and a GVF-Snake algorithm were applied to estimate the LV contour. This pipeline of processes was performed without any operator intervention and its lower performance is due to the possible failure of one of the steps involved in the whole process.

Despite of these good estimates, semi and largely automated methods were not ranked as well as the manual approaches.

Their actual version should be improved to be clinically accepted.

B. Robustness and Limits of the eRWT Approach

When modifying the Beta distribution parameters, we observed that the figures of merit for each segmentation method were the smallest when the Beta distribution parameters were set *a priori* and not when they were estimated during the eRWT process. Moreover, we found that the smallest F_m were obtained when the Beta distribution parameters μ and ν described a database with more reduced EFs than normal values (cf. Fig. 6). In addition, with these distributions, manual delineations were ranked first whereas it was not the case with the other distributions (cf. Table II). These observations showed the importance to carefully describe *a priori* the distribution of the clinical parameter of interest for a given database. This also showed that the computation of the figure of merit can help to choose the most appropriate set of parameters of the Beta distribution.

Comparisons performed when modifying the number of segmentation methods showed that the eRWT approach accurately and reproducibly ranked the methods for any combination of at least three methods (cf. Fig. 7). Remaining errors were characterized with a ranking inversion cost less than 2, which means a possible error of one rank only. We also observed that eRWT results might not be reliable when only two methods are compared.

When comparing the methods on the reduced database (thirty patients), the overall ranking of the eight segmentation methods was close to the one obtained on the whole database (cf. Section V-B3). However, when only three methods entered the comparison, the percentage of changes in the methods ranking was a little bit higher with the reduced database than when the whole database was included (cf. Fig. 7). This showed that, even if a bootstrap approach is used on top of the comparison process, reducing the size of the database might decrease the probability to correctly rank three methods. Larger database should be preferred when only three segmentation methods have to be compared to each other.

C. Future Directions

A strong assumption of the RWT techniques is that the estimated values are linearly related to the true values [4], [7], [8]. Results presented in these previous reports and our findings suggest that such an assumption is reasonable in a large number of cases. Yet, the use of a quadratic model has actually been proposed in [22]. The drawback of such an extension is that an additional parameter has to be estimated for each method. It is therefore more robust to use the linear assumption when it properly describes the data. Results shown in Fig. 4 suggest that this is the case in our dataset. However, additional studies should be carried out in the future to determine whether the linear assumption is reasonable for a given dataset, or whether a different assumption should be used instead.

Our results were obtained using a constrained nonlinear multivariable function based on a sequential quadratic programming method [19]. To ensure convergence of the algorithm, the initialization of the algorithm was set close to the expected solution. Additional tests will be performed to study the influence of the initialization on the results. Other algorithms, based on a stochastic clustering, gradient descent and simulated annealing, will also be tested.

Our eRWT method allows to conclude at the statistical significance of differences in the ranking of estimation methods. Yet, it does not address the practical significance of differences in ranking. To go in greater depth in the methods ranking, the eRWT approach could be applied to analyze other relevant clinical parameters, like the end-diastolic and end-systolic volumes. If several methods have similar performance according to the eRWT approach, additional criteria, like the computation time required to provide the segmentation results, could also be taken into account to conclude that a method is more practical than another.

When considering the performance level of image segmentation approaches without gold standard, another solution consists in evaluating the accuracy of the segmentation results by estimating a reference shape from the segmentation entries. Using this estimate, some associated evaluation parameters can be computed (e.g., sensibility, specificity, distance to the estimated reference shape). Among the methods that will be investigated, the now classical STAPLE algorithm [23] and another method based on variational approaches and active contours [24] will be tested. In this latter method, we propose to estimate a mutual reference shape within a continuous optimization framework by minimizing a criterion based on information theory. Let us note that the contour-based methods cited above provide additional information on the relevance of each segmentation method. They may then also be used jointly with the eRWT approach in order to design a more complete evaluation framework.

VII. Conclusion

The present study demonstrates the robustness and the limitations of an extended version of the RWT approach for comparing the accuracy of different segmentation methods used to estimate, for a database, a clinically relevant parameter in the absence of gold standard (here the ejection fraction in cardiac MRI). In comparison with previous applications, this extended methodology was applied to rank numerous methods (eight in total). No *prior* concerning the reliability or the degree of automation of the segmentation methods is required to perform this comparison without a gold standard. A few conditions must nevertheless be respected to apply the methodology. First, a *prior* concerning the distribution of the biomarker is highly recommended. Additionally, our results suggest that eRWT provides an accurate ranking of methods when the database includes at least 30 samples and when at least three methods are compared. Manual delineation might therefore not be required anymore to evaluate the relative performance of different segmentation algorithms.

Acknowledgment

The authors would like to thank A. Cansot for his help in performing the robustness tests and C. Pellot-Barakat for her careful reading.

REFERENCES

[1] D. P. Huttenlocher, G. A. Klanderman, and W. A. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, pp. 850–863, Sep. 1993.

[2] A. Rosenfeld and J. L. Pfaltz, "Sequential operations in digital picture processing," *J. ACM*, vol. 13, pp. 471–494, Oct. 1966.

[3] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, pp. 297–302, Jul. 1945.

[4] M. A. Kupinski, J. W. Hoppin, J. Krasnow, S. Dahlberg, J. A. Leppo, M. A. King, E. Clarkson, and H. H. Barrett, "Comparing cardiac ejection fraction estimation algorithms without a gold standard," *Acad. Radiol.*, vol. 13, pp. 329–337, Mar. 2006.

[5] K. R. Choudhury, D. S. Paik, C. A. Yi, S. Napel, J. Roos, and G. D. Rubin, "Assessing operating characteristics of CAD algorithms in the absence of a gold standard," *Med. Phys.*, vol. 37, pp. 1788–1795, Apr. 2010.

[6] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 1, pp. 307–310, Feb. 1986.

[7] J. W. Hoppin, M. A. Kupinski, G. A. Kastis, E. Clarkson, and H. H. Barrett, "Objective comparison of quantitative imaging modalities without the use of a gold standard," *IEEE Trans. Med. Imag.*, vol. 21, no. 5, pp. 441–449, May 2002.

[8] M. A. Kupinski, J. W. Hoppin, E. Clarkson, H. H. Barrett, and G. A. Kastis, "Estimation in medical imaging without a gold standard," *Acad. Radiol.*, vol. 9, pp. 290–297, Mar. 2002.

[9] A. K. Jha, M. A. Kupinski, J. J. Rodríguez, R. M. Stephen, and A. T. Stopeck, "Evaluating segmentation algorithms for diffusion-weighted MR images: A task-based approach," *Proc. Soc. Photo. Opt. Instrum. Eng.*, vol. 7627, pp. 76270L1–76270L8, Feb. 2010.

[10] M. Soret, J. Alaoui, P. M. Koulibaly, J. Darcourt, and I. Buvat, "Accuracy of partial volume effect correction in clinical molecular imaging of dopamine transporter using SPECT," *Nucl. Instrum. Meth. Phys. Res. Sec. A*, vol. 571, pp. 173–176, Feb. 2007.

[11] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman Hall, 1993.

[12] J. Lebenberg, I. Buvat, M. Garreau, C. Casta, C. Constantinidès, J. Cousty, A. Cochet, S. Jehan-Besson, C. Tilmant, M. Lefort, E. Roullot, L. Najman, L. Sarry, P. Clarysse, A. De Cesare, A. Lalande, and F. Frouin, "Comparison of different segmentation approaches without using gold standard. Application to the estimation of the left ventricle ejection fraction from cardiac cine MRI sequences," in *Proc. 33rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Boston, MA, Aug.–Sep. 30–4, 2011, pp. 2663–2666.

[13] P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick, and G. Wright, "Evaluation framework for algorithms segmenting short axis cardiac MRI," *MIDAS J.—Cardiac MR Left Ventricle Segmentation Challenge*, 2009.

[14] C. Constantinides, Y. Chenoune, N. Kachenoura, E. Roullot, E. Mousseaux, A. Herment, and F. Frouin, "Semi-automated cardiac segmentation on cine magnetic resonance images using GVF-Snake deformable models," *MIDAS J.—Cardiac MR Left Ventricle Segmentation Challenge*, 2009.

[15] J. Schaerer, C. Casta, J. Pousin, and P. Clarysse, "A dynamic elastic model for segmentation and tracking of the heart in MR image sequences," *Med. Imag. Anal.*, vol. 14, pp. 738–749, Dec. 2010.

[16] J. Cousty, L. Najman, M. Couprie, S. Clement-Guinaudeau, T. Goissen, and J. Garot, "Segmentation of 4D cardiac MRI: Automated method based on spatio-temporal watershed cuts," *Image Vis. Comput.*, vol. 28, pp. 1229–1243, Aug. 2010.

[17] A. Lalande, N. Salvé, A. Comte, M. C. Jaulent, L. Legrand, P. M. Walker, Y. Cottin, J. E. Wolf, and F. Brunotte, "Left ventricular ejection fraction calculation from automatically selected and processed diastolic and systolic frames in short-axis cine-MRI," *J. Cardiovasc. Magn. Reson.*, vol. 6, pp. 817–827, 2004.

[18] C. Constantinides, Y. Chenoune, E. Mousseaux, F. Frouin, and E. Roullot, "Automated heart localization for the segmentation of the ventricular cavities on cine magnetic resonance images," in *Proc. Computing in Cardiology*, Belfast, U.K., Sep. 26–29, 2010, vol. 37, pp. 911–914.

[19] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*. London, U.K.: Academic, 1981.

[20] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *J. Amer. Stat. Assoc.*, vol. 47, pp. 583–621, Dec. 1952.

[21] R. G. Miller, *Simultaneous Statistical Inference*, 2nd ed. New York: Springer Verlag, 1981.

[22] I. Buvat, J. Alaoui, and M. Soret, "Comparison of estimation methods without a gold standard," *J. Nucl. Med.*, vol. 47, pp. 116P-a–116P-a, May 2006.

[23] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.

[24] S. Jehan-Besson, C. Tilmant, A. De Cesare, F. Frouin, L. Najman, A. Lalande, L. Sarry, C. Casta, P. Clarysse, C. Constantinidès, J. Cousty, M. Lefort, A. Cochet, and M. Garreau, "Estimation d'une forme mutuelle pour l'évaluation de la segmentation en imagerie cardiaque," in *XXIIIème Colloque du Groupe d'Etudes du Traitement du Signal et des Images (GRETSI)* (in French), Bordeaux, France, Sep. 5–8, 2011.