Medical Physics
The International Journal of Medical Physics Research and Practice

A framework for quantification and visualization of segmentation accuracy and variability in 3D lateral ventricle ultrasound images of preterm neonates

Yimin Chen, Wu Qiu, Jessica Kishimoto, Yuan Gao, Rosa H. M. Chan, Sandrine de Ribaupierre, Aaron Fenster, and Bernard Chiu

Citation: Medical Physics **42**, 6387 (2015); doi: 10.1118/1.4932366
View online: http://dx.doi.org/10.1118/1.4932366
View Table of Contents: http://scitation.aip.org/content/aapm/journal/medphys/42/11?ver=pdfcov
Published by the American Association of Physicists in Medicine

A framework for quantification and visualization of segmentation accuracy and variability in 3D lateral ventricle ultrasound images of preterm neonates

Yimin Chen[a)]
Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong

Wu Qiu[a)] and Jessica Kishimoto
Imaging Research Laboratories, Robarts Research Institute, The University of Western Ontario, London, Ontario N6A 5K8, Canada

Yuan Gao and Rosa H. M. Chan
Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong

Sandrine de Ribaupierre
Department of Clinical Neurological Science, The University of Western Ontario, London, Ontario N6A 5K8, Canada

Aaron Fenster
Imaging Research Laboratories, Robarts Research Institute, The University of Western Ontario, London, Ontario N6A 5K8, Canada

Bernard Chiu[b)]
Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong

(Received 1 April 2015; revised 13 August 2015; accepted for publication 22 September 2015; published 13 October 2015)

Purpose: Intraventricular hemorrhage (IVH) is a major cause of brain injury in preterm neonates. Three dimensional ultrasound (US) imaging systems have been developed to visualize 3D anatomical structure of preterm neonatal intracranial ventricular system with IVH and ventricular dilation. To allow quantitative analysis, the ventricle system is required to be segmented accurately and efficiently from 3D US images. Although semiautomatic segmentation algorithms have been developed, local segmentation accuracy and variability associated with these algorithms should be evaluated statistically before they can be applied in clinical settings. This work proposes a statistical framework to quantify the local accuracy and variability and performs statistical tests to identify locations where the semiautomatically segmented surfaces are significantly different from manually segmented surfaces.
Methods: Three dimensional lateral ventricle US images of preterm neonates were each segmented six times manually and using a semiautomated segmentation algorithm. The local difference between manually and algorithmically segmented surfaces as well as the segmentation variability for each method was computed and superimposed on the ventricular surface of each subject. To summarize the segmentation performance for a whole group of subjects, the subject-specific local difference and standard deviation maps were registered onto a 3D template ventricular surface using a nonrigid registration algorithm. Pointwise, intersubject average accuracy and pooled variability for the whole group of subjects can be computed and visualized on the template surface, providing a summary of performance of the segmentation algorithm for the whole group of ventricles with highly variable geometry. In addition to pointwise statistical analysis performed on the template surface, statistical conclusion regarding the accuracy of the segmentation algorithm was made for subregions and the whole ventricle with the spatial correlation of pointwise accuracy taken into account.
Results: Ten 3D US images were involved in this study. Pointwise local difference, ΔS, its absolute value $|\Delta S|$ as well as the standard deviations of the manual and algorithm segmentations were computed and superimposed on the each ventricle surface. Regions with lower segmentation accuracy and higher segmentation variability can be identified from these maps, and the localized information was applied to improve the accuracy of the algorithm. Intersubject average ΔS and $|\Delta S|$ as well as pooled standard deviations was computed on the template surface. Intersubject average ΔS and $|\Delta S|$ indicated that the algorithm underestimated regions in the neighborhood of the tips of anterior, inferior, and posterior horns. Intersubject pooled standard deviations indicated that manual segmentation had a higher segmentation variability than algorithm segmentation over the whole ventricle. Statistical analysis on the template surface showed that there was significant difference between algorithm and manual methods for segmenting the right lateral ventricle but not for the left lateral ventricle.
Conclusions: A framework was proposed for evaluating, visualizing, and summarizing the local accuracy and variability of a segmentation algorithm. This framework can be used for improving the accuracy of segmentation algorithms, as well as providing useful feedback to improve the manual

segmentation performance. More importantly, this framework can be applied for longitudinal monitoring of local ventricular changes of neonates with IVH. © *2015 American Association of Physicists in Medicine.* [http://dx.doi.org/10.1118/1.4932366]

Key words: segmentation accuracy, preterm neonate, cerebral ventricle, intraventricular hemorrhage (IVH), 3D ultrasound

1. INTRODUCTION

Among very low birth weight preterm neonates, the most common noncongenital cause of cerebral ventricle dilation is intraventricular hemorrhage (IVH).[1] Post hemorrhagic ventricular dilatation (PHVD) is linked to specific neurolopsychological impairments, such as visuospatial and motor deficits.[2] Interventions, such as ventricle tapping and "shunt" insertion, have been introduced to improve neurolopsychological outcomes. Interventional decisions are typically made based on clinical signs such as increased head circumference, bulging of the anterior fontanelle, along with qualitative, visually based increased ventricle size from ultrasound (US) images as well as symptoms of increased intracranial pressure (ICP), which tend to be nonspecific in this patient population. Therefore, monitoring neonatal ventricles in a more quantitative manner may allow clinicians to make better interventional decisions.

MRI has been established as the golden standard for ventricular volume measurements.[3,4] However, this technique is expensive and time consuming, with the child sometimes needing to be sedated to avoid motion artifacts. Due to its cost-effectiveness, 2D cranial US imaging is performed as standard care on preterm neonates to diagnose and monitor IVH and the subsequent ventricular dilation. Two dimensional US has shown good correlation with MRI in measurements of frontal horns and ventricular midbody;[5,6] however, 2D US measurements are prone to operator and interscan variability due to the requirement of selecting a "slice" of the ventricle to image. To address this issue, 3D ultrasound systems have been developed to obtain more accurate and reproducible measurements of the ventricle volume[7,8] at the bed side.

Volumetric measurement of cerebral ventricle requires manual segmentation from the 3D US image, which is laborious and time-consuming. In a previous study,[8] 20–30 min were required to manually segment ventricles from a subject US image. In addition, manual segmentation is prone to variability due to speckle, shadowing, and noise. In order to reduce human interactions and user variability in ventricle segmentation, a semiautomatic segmentation algorithm has been proposed to segment preterm neonate ventricle from 3D US images.[9] The algorithm was validated by comparing with manual segmentation, which is widely used as the surrogate gold standard in segmentation evaluation.[9–11] The algorithm was able to obtain ventricular volume measurements that were not statistically different from manual measurements. The coefficient of variation in ventricular volume measurements obtained from repeated algorithm segmentations was also shown to be smaller than that obtained from repeated manual segmentations,[9] which is not unexpected as human interactions, the source of variability, were limited to the initialization stage in algorithm segmentation. Volumetric metrics, such as ventricular volume and volume overlap, although widely used in segmentation evaluation,[10,11] do not provide information as to the regions within the ventricle associated with large error and variability. Localizing the regions with greater disagreement between algorithm and manual segmentations would allow development of strategies for improving a semiautomated segmentation algorithm, such as increasing the number of initial points locally within regions where segmentation accuracy is required to be improved.

Although in the brain morphometry field, registration and statistical models have been proposed to quantify local anatomical changes observed primarily from MRI,[12–14] segmentation evaluation in 3D neonatal ventricle ultrasound images poses unique challenges that could not be addressed by existing frameworks. One major component of brain morphometry framework is to register individual MR images to a template brain image in order to compute a deformation field for later local statistical analysis. In the brain morphometry literature,[15,16] MR images acquired from different subjects in the study group were somewhat similar, which is not the case for 3D ultrasound images of neonatal ventricles with IVH. As demonstrated in Fig. 2 in Qiu *et al.*,[9] neonatal ventricles with IVH have irregular shapes, which are also highly subject-specific. Image quality issues such as intensity inhomogeneity inside the ventricle, missing edges, and hyperechogenicity due to IVH would all preclude accurate image registration. Surface correspondence algorithm built in to some morphometry pipelines also requires surfaces to be matched to have similar shapes. In developing the surface correspondence algorithm for matching cortical surfaces acquired in a longitudinal MRI study, Chung *et al.*[17] assumed cortical surface displacement to be less than 1 mm/yr on average. Due to image quality issues of 3D ultrasound images of neonatal ventricles described above, local difference between algorithm and manual segmentations can be up to 8 mm. Clearly, a more robust surface correspondence algorithm is required for the development of our segmentation evaluation framework.

In Mao *et al.*[18] and Chiu *et al.*,[19] boundaries were segmented in transverse slices. Pointwise evaluation of the segmentation algorithm was based on matching 2D contours without considering 3D geometry of the segmented boundaries. The evaluation method applied in Gill *et al.*[20] is 3D surface-based, in which each vertex on the algorithm segmented surface corresponds to the closest point on the manually segmented surface. This approach for establishing correspondence is not symmetric (i.e., the correspondence relationship would be different if the roles of the two surfaces were interchanged). Local segmentation distance error depends on whether we are finding the closest point on the manually segmented surface

for each vertex on the algorithm segmented surface or the other way around. Therefore, a metric for quantifying the local distance between the manual and algorithm segmentations could not be uniquely defined. Moreover, the local distance would be underestimated in bulgy surfaces by this approach. In this paper, we propose a multiresolution approach to establish 3D symmetric correspondence between the manually and algorithmically segmented surfaces in order to allow for local segmentation accuracy and variability evaluation.

Moreover, as there is a requirement for the proposed statistical framework to consider observer variability estimated from repeated segmentations, the statistical framework developed here must accommodate one more level of complexity than existing morphometry pipelines. A major application of brain morphometry pipelines, such as the deformation-based surface morphometry,[17] is to make statistical conclusion on whether there is a difference in a metric (e.g., cortical thickness and curvature) between two groups of subjects at a specific position of the brain template. The conclusion can be made based on the mean difference of the metric in relation to the standard error derived from intersubject variability. In comparison, our framework was designed to make statistical conclusions at each point on the ventricle surface on whether the algorithm segmentation is significantly different from the surrogate ground truth (i.e., manual segmentation) on two levels: the subject and the group levels. On the subject level, the statistical conclusion was made in relation to the segmentation variability estimated from repeated segmentations at each vertex. The proposed framework is well equipped to (i) generate a mean surface from repeated segmentations, (ii) establish a correspondence relationship between the algorithm and manual segmentations so that a pointwise distance between them can be measured, and (iii) estimate segmentation variability at each vertex based on repeated segmentations. On the group level, the statistical conclusion on the difference between algorithm and manual segmentation was drawn on each point of the ventricle template. At this level, the framework should be capable of mapping each individual ventricle to the ventricle template and, at each point on the template, assessing the mean difference over the whole population of subjects in relation to the segmentation variabilities estimated at the subject level. Although many methods have been proposed for registering adult brain ventricle segmented from MR images,[21–24] these methods are not suitable for neonatal ventricles with IVH segmented from 3D US images because of the irregular shape deformation caused by IVH and degraded image quality. This paper describes the development of a nonrigid shape registration method, which we validated to be robust in mapping individual ventricular surfaces of a population of neonates with IVH to a standard ventricular template.

In summary, the goal of this paper is to develop a statistical framework to allow for the evaluation of a semiautomated algorithm designed for neonatal ventricle segmentation from 3D ultrasound images. The novelty of this framework is that statistical conclusion regarding accuracy of the segmented surfaces can be made on the subject and group levels. The major challenges include (i) the development of surface correspondence algorithm to match algorithm and manual segmentations, (ii) estimation of segmentation variability from repeated segmentations, (iii) registration of individual ventricle to a ventricle template, and (iv) derivation of statistics that allow conclusions to be drawn on the accuracy of the segmentation algorithm on the subject and group levels. Since the segmentation accuracy and variability metrics as well as the statistical conclusion based on these metrics are available on a point-by-point basis, and also on the subject and group levels, these quantities can be visualized as distributions superimposed on the individual or template ventricle surfaces. These distributions allow for the identification of positions where improvement of segmentation results is required. Although so far, we have focused our discussion on statistical tests performed on each point of the ventricle; these tests were extended to regional and global scales (i.e., statistical conclusions are made for a region of or the whole ventricle) that take into account of spatial correlation of segmentation accuracy metrics between neighboring points. In this paper, we also demonstrate results for these regional and global statistical analyses.

2. MATERIALS AND METHODS

2.A. Study subjects and ultrasound image acquisition

In this study, ten 3D ultrasound images were segmented for both left and right ventricles. Subjects involved in this study represent a wide range of IVH severity: (1) three patients with IVH III; (2) two patients with IVH II; and (3) five patients with IVH I. All patients provided written informed consent to the study protocol, which was approved by the ethics review board at The University of Western Ontario. These patients were diagnosed of IVH initially with clinical 2D ultrasound exam and were subsequently imaged with 3D ultrasound.

3D ultrasound images were acquired with a motorized 3D ultrasound system developed for cranial scanning of preterm neonates in Neonatal Intensive Care Unit in University Hospital of The University of Western Ontario, which used an HDI 5000 (Philips, Bothell, WA) and C8-5 (Philips, Bothell, WA) curved array 5–8 MHz broadband transducer.[8] To perform a scan, an ultrasound technician located the third ventricle, midline through the anterior fontanelle with the patient inside an incubator, and then the 3D ultrasound system mechanically tilted the 2D transducer to acquire a full 3D image of the ventricular system. The 3D image sizes ranged from $300\times300\times300$ to $450\times450\times450$ voxels with a voxel spacing of $0.22\times0.22\times0.22$ mm^3.

2.B. Manual and algorithm segmentations

Six repeated segmentations produced by each of the manual and algorithm segmentation methods were analyzed and compared. These two segmentation methods are briefly summarized below.

3D ultrasound images were loaded into an in-housed analysis software and displayed using a multiplanar texture mapping approach.[25] Two trained observers segmented each subject image in parallel slices with 1 mm interval through sagittal

planes as shown in Fig. 2(a). This process was repeated three times for each observer with a 24-hr interval between consecutive segmentation sessions with the 10 3D ultrasound images randomized in each session. Each observer was blinded to the image order to reduce memory bias.

The algorithm segmentation approach required users to manually label several voxels inside and outside ventricles as foreground and background on a few sagittal views. Each voxel in a 3D ultrasound image was then classified into foreground and background using a convex optimization algorithm as previously described in Qiu *et al.*[9] In this study, two trained observers repeated the initialization procedure for three times for each image with a 24-hr interval and six algorithm segmentations were obtained for each image.

2.C. Computation of mean surfaces for repeated manual and algorithm segmentations

Figure 1 shows the schematic diagram describing the steps required for generating the mean surfaces from repeated manual and algorithm segmentations. In this study, manual and algorithm segmentations were each repeated six times. For both segmentation methods, a 3D indicator function with 1 and 0 indicating the foreground and background, respectively, represents each of the repeated segmentations. For each of the manual and algorithm segmentation methods, six 3D indicator functions representing repeated segmentations served as inputs to a previously described probabilistic framework called Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm,[26] which generated an output 3D indicator function representing the segmentation estimated from repeated segmentations. The marching cube algorithm[27] was subsequently used to convert this indicator function to a 3D surface, which we refer to as mean surface hereafter. The mean surfaces generated for the repeated segmentations produced manually and by the algorithm will be used in the definition of local accuracy and variability metrics.

For the algorithm segmentation, the 3D indicator function was available as the output of the algorithm as described previously. However, as manual segmentation was performed on a slice-by-slice basis, there is a need to generate a 3D indicator function for each stack of manually segmented boundaries and this was achieved as follows. For each contour segmented from each transverse slice of the 3D ultrasound image, a distance map was generated indicating the shortest signed distance from each point on the image to the contour as described previously[28,29] and shown in Fig. 2(b). Points inside and outside the contour were equipped with positive and negative values, respectively. Distance values at points between two adjacent transverse images were obtained by linearly interpolating distance maps on the transverse images, resulting in a 3D distance map. A marching cube algorithm was applied to generate a 3D surface by extracting the zero level set of the 3D distance map [Fig. 2(c)]. The 3D indicator function representing the manually segmented boundary was generated with voxel inside the surface assigned a gray level of 1 and the remaining voxels assigned a gray level of 0 [Fig. 2(d)]. This process was performed for each of the six stacks of repeated manual segmentations, resulting in six 3D indicator functions used for estimating the mean surface as described in the previous paragraph.

2.D. Statistical inferences on the differences between manual and algorithm segmentations

2.D.1. Pointwise comparison

The matching of the mean manual and algorithm segmentations generated in Sec. 2.C is required before computing the local difference between these two surfaces. Papademetris *et al.*[30] proposed a symmetric correspondence algorithm to find a point-by-point correspondence based on symmetric nearest points. This algorithm consists of two major steps: (a) identification of symmetric correspondence pairs between two curves

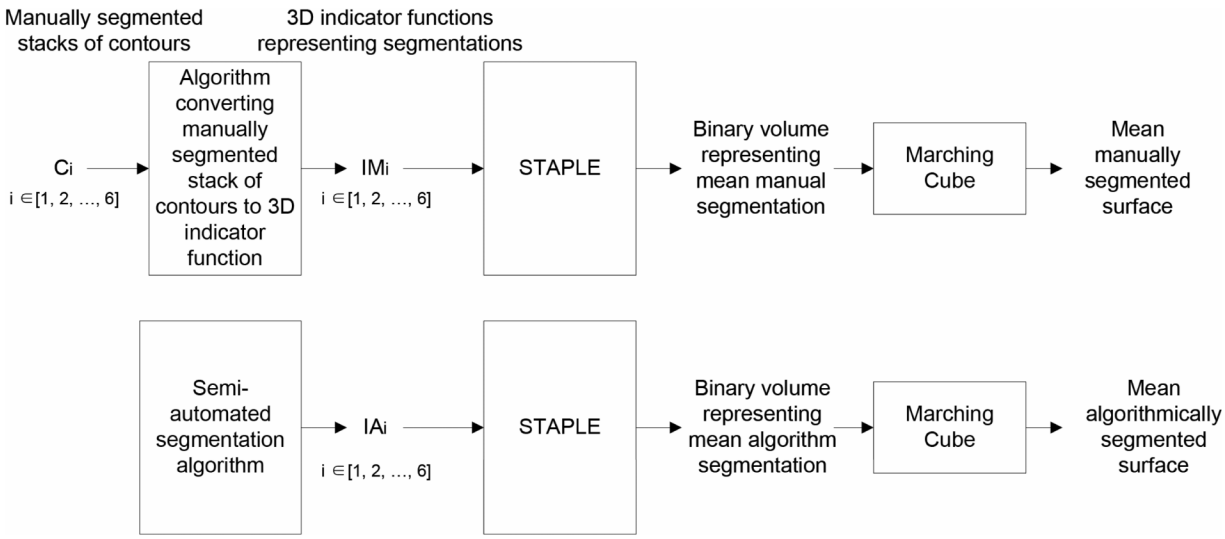

FIG. 1. Schematic diagram of generation of mean surfaces for manual and algorithm, respectively. C_i $(i = 1, 2, \ldots, 6)$ denotes a series of manual contours for each manual segmentation. IM_i and IA_i $(i = 1, 2, \ldots, 6)$ are the 3D indicator functions for manual and algorithm segmentations, respectively.

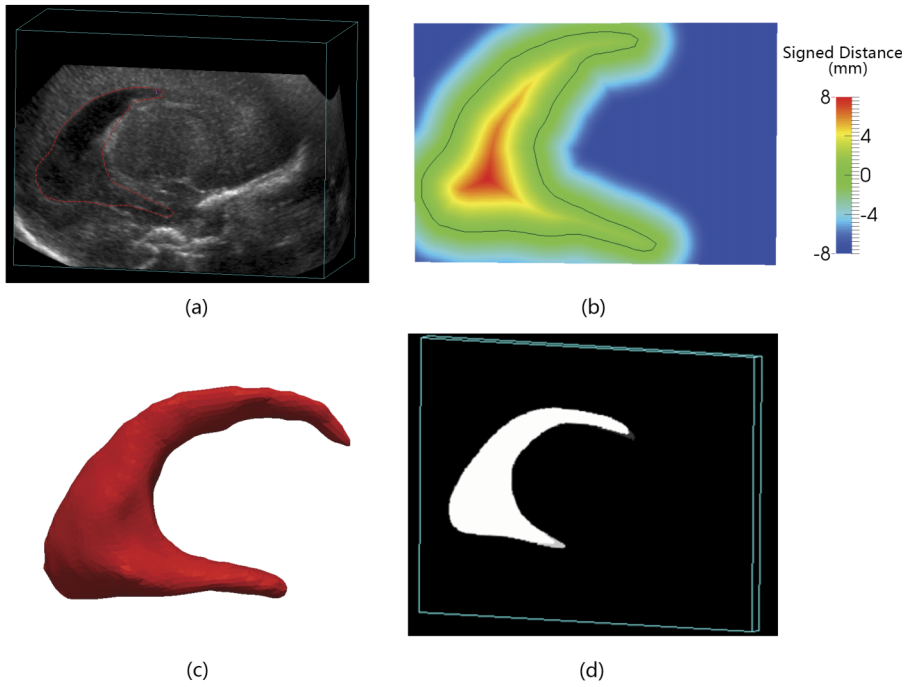

FIG. 2. (a) A manually segmented contour in one slice of a 3D ultrasound image. (b) A signed distance map showing the distance from each voxel to the boundary. (c) 3D surface generated by extracting the zero-level set of the distance map. (d) A 3D binary volume with voxels inside the segmented surface shown in (c) assigned 1 and the remaining voxels assigned 0.

(or surfaces in 3D) and (b) interpolation to match segments between two symmetric correspondence pairs. For example, four pairs of correspondence were labeled in Fig. 3(a) for two curves and the correspondence pairs were numbered according to the clockwise orientation of C_1. The interpolation step requires segments between any 2 pairs of symmetric correspondence in both curves be nonoverlapping. This requirement was violated for C_2 in which the segment 2–3 and segment 3–4 overlap. Consequently, interpolation cannot map C_2 to C_1 on a one-to-one basis. The segment between pairs 3 and 4 in C_2 can either be mapped to the segment between pairs 2 and 3 or that between pairs 3 and 4 in C_1. This problem occurs largely because of the existence of intersections between the two curves being matched.

In this paper, we address this problem by introducing a multiresolution approach in which ventricle surfaces to be matched were first divided based on their intersection into patches [analogous to segments in 2D shown in Fig. 3(b)], which were then matched according to Boolean operations on surfaces as described in Quammen *et al.*[31] based on two criteria: (1) the corresponding patches must have the same boundary. (2) Let the two surfaces to be matched be $\mathcal{A}$ and $\mathcal{B}$. A patch on surface $\mathcal{A}$ outside surface $\mathcal{B}$ must be matched to a patch on surface $\mathcal{B}$ that is inside surface $\mathcal{A}$. Similarly, the patch on surface $\mathcal{A}$ inside surface $\mathcal{B}$ must be matched to a patch on surface $\mathcal{B}$ that is outside surface $\mathcal{A}$. Figure 4 shows an example of the patch correspondence operation. Figure 4(b) shows the intersection lines between the two surfaces shown in Fig. 4(a). Corresponding patches are displayed in the same color in Figs. 4(c) and 4(d). Each pair of matched patches was then individually matched pointwise using Papademetris's 3D symmetric correspondence algorithm.

For each point (p_M) on manual mean surface ($\overline{S_M}$), we could find its corresponding point (p_A) on mean algorithm surface ($\overline{S_A}$) with the combination of our multiresolution approach and 3D symmetric correspondence algorithm. The

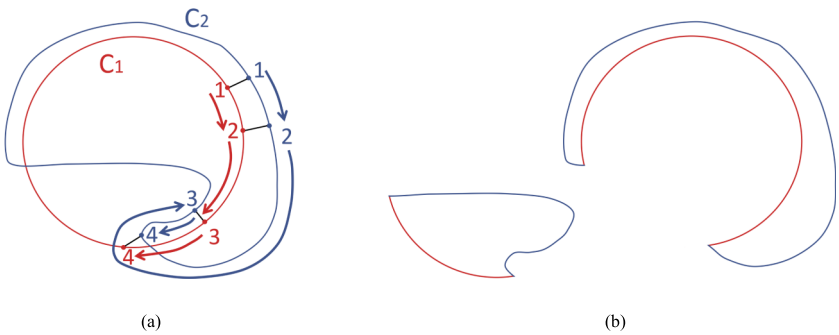

FIG. 3. (a) Symmetric correspondence pairs for two curves with Papademetris *et al.* (Ref. 30) and (b) the curves are decomposed into patches with their intersections.

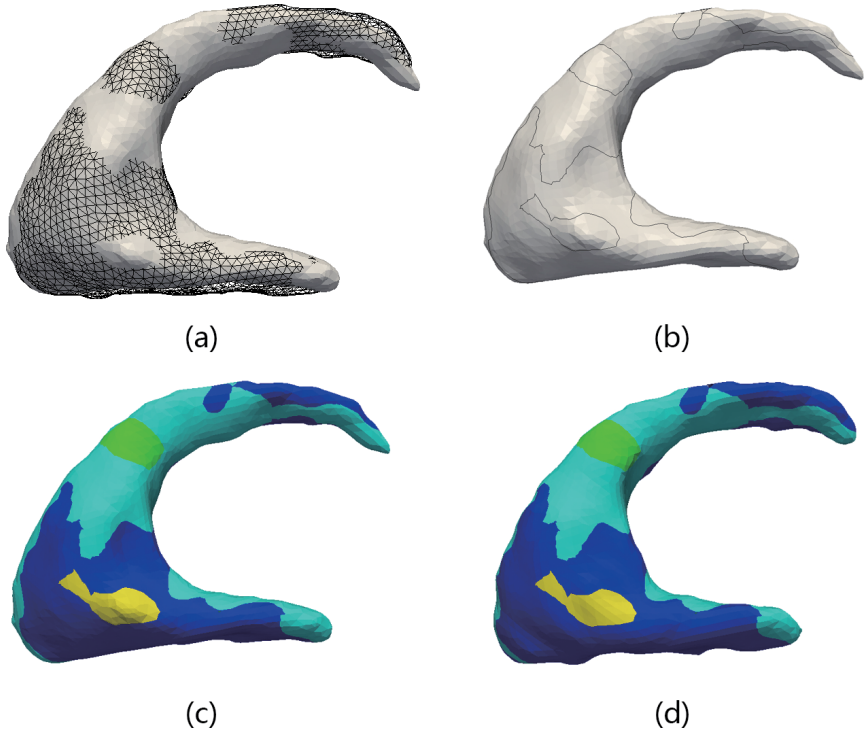

FIG. 4. Patch correspondence between two surfaces. (a) shows two mean surfaces overlapped, and (b) shows the intersection lines (black lines) between those two mean surfaces superimposed on the manual surface [solid surface in (a)]. (c) and (d) show the patches generated from the intersection lines of those two surfaces and pairs of corresponding patches are displayed in the same color.

distance between them is defined as local difference (ΔS) between p_M and p_A,

$$\Delta S = \begin{cases} \|p_M - p_A\|, & \text{if } p_A \text{ is outside of } \overline{S_M} \\ -\|p_M - p_A\|, & \text{if } p_A \text{ is inside of } \overline{S_M} \end{cases}. \quad (1)$$

A pair of points (p_M and p_A) defines a line that intersects with six manual and algorithm surfaces. The distance between each of the six manual intersections and p_M can be obtained. The standard deviation of this group of six distances is defined as the local standard deviation of manual segmentation (SD_M). Similar method was applied to compute the local standard deviation of algorithm segmentation, denoted as SD_A. The standard error of the mean distance between surfaces segmented by these two methods was defined as

$$\mathrm{SE}_{\Delta S} = \sqrt{\frac{\mathrm{SD}_A{}^2}{n_A} + \frac{\mathrm{SD}_M{}^2}{n_M}}, \quad (2)$$

where n_A and n_M represent the number of surfaces from algorithm and manual, respectively, that have an intersection with the line defined by two points p_A and p_M. Point-by-point T-tests are performed in order to evaluate where the mean distance is significantly different from 0. The T-statistics and the degree of freedom are given by

$$T = \frac{\Delta S}{\mathrm{SE}_{\Delta S}}, \quad (3)$$

$$\nu = \frac{\left(\frac{\mathrm{SD}_A{}^2}{n_A} + \frac{\mathrm{SD}_M{}^2}{n_M}\right)^2}{\frac{\left(\mathrm{SD}_A{}^2/n_A\right)^2}{n_A - 1} + \frac{\left(\mathrm{SD}_M{}^2/n_M\right)^2}{n_M - 1}}. \quad (4)$$

2.D.2. Regional comparison

Although the pointwise statistical test introduced in Eqs. (3) and (4) provided rich information on the locations where statistically significant differences occur between the algorithm and manual segmentations, the test treats ΔS at each vertex as an independent quantity and does not take the spatial correlation of ΔS into consideration. In addition, a statistical test on ΔS over a region of interest (ROI) is useful for quantitative characterization of the performance of the segmentation algorithm in relation to the quality of the 3D ultrasound images (e.g., missing edges, intensity inhomogeneity, and irregular shape due to IVH at the inferior and posterior horns described in Qiu *et al.*[9]). Although clinicians would still be able to perform qualitative regional analysis by visualizing the pointwise 3D T-map generated by Eqs. (3) and (4) [e.g., Fig. 7(e)], a ROI-based statistical test would make a quantified regional analysis possible without the introduction of observer variability in the process.

ΔS is modeled as a Gaussian random field on a ROI (Ω) of the ventricle surface as described in Chung *et al.*[17] Under the null hypotheses (i.e., no difference between manual and algorithm segmentation in Ω),

$$H_0 : \Delta S = 0 \text{ for all } p \in \Omega. \quad (5)$$

The pointwise T-statistic at a vertex p on the ventricle surface, denoted by $T(p)$ and given by Eq. (3), is T-distributed with ν degrees of freedom given by Eq. (4). The supremum of T values over the ROI Ω, denoted by $\sup_{p\in\Omega} T(p)$, has the following distribution:[17]

$$P\left(\sup_{p\in\Omega} T(p) > h\right) \approx \sum_{i=0}^{2} \phi_i(\Omega)\rho_i(h), \tag{6}$$

where h is a constant, ρ_i is the ith dimensional density of Euler characteristic, and ϕ_i is the Minkowski functional of Ω, which depends on the topology of Ω. In this study, we performed this test for selected regions of the ventricle as well as for the whole ventricle surface. For the first case, the Minkowski functionals are expressed by $\phi_0(\Omega) = 1$, $\phi_1(\Omega) = L/2$, and $\phi_2(\Omega) = \|\Omega\|$, where L is the perimeter of the selected region and $\|\Omega\|$ is the surface area of the region. For the second case, since the whole ventricle surface is a closed surface, we have $\phi_0(\Omega) = 2$, $\phi_1(\Omega) = 0$, and $\phi_2(\Omega) = \|\Omega\|$. In this statistical framework, the correlation between ΔS at neighboring points was quantified by treating ΔS as a Gaussian random field with full-width-half-maximum (FWHM) given by[32]

$$\mathrm{FWHM} = \overline{\mathrm{edge}}\sqrt{\frac{-2\ln 2}{\ln\left(1-\frac{\mathrm{var}(d\Delta S)}{\mathrm{var}(\Delta S)}\right)}}, \tag{7}$$

where $\overline{\mathrm{edge}}$ is the average interneighbor distance, $\mathrm{var}(d\Delta S)$ is the variance interneighbor differences of ΔS value, and $\mathrm{var}(\Delta S)$ is the variance of ΔS computed over Ω.

The Minkowski functionals are given by[33]

$$\rho_0(h) = \int_0^\infty \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{(\nu\pi)^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-((\nu+1)/2)} dx, \tag{8}$$

$$\rho_1(h) = \frac{\lambda^{1/2}}{2\pi}\left(1+\frac{h^2}{\nu}\right)^{-1/2(\nu-1)}, \tag{9}$$

$$\rho_2(h) = \frac{\lambda}{(2\pi)^{3/2}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\left(\frac{\nu}{2}\right)^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{h^2}{\nu}\right)^{-((\nu-1)/2)}, \tag{10}$$

where $\lambda = 4\ln 2/\mathrm{FWHM}^2$, Γ is the gamma function, and ν is the degrees of freedom for the T-statistics.

Notably, the number of degrees of freedom ν of the T-statistic varies among points since it depends on the ratio between SD_A^2 and SD_M^2 as given by Eq. (4). The statistical distribution of $\sup_{p\in\Omega} T(p)$ was derived based on the assumption that ν at all points is equal. In this study, we pooled the SD_A^2 and SD_M^2 over all subjects (Table II). The pooled segmentation variabilities were used to calculate ν.

2.E. Evaluation of segmentation methods for a whole group of subjects

In order to reach a conclusion applicable to the whole group of subjects, all surfaces must be mapped to a standard template shape, in which the local evaluation metrics for different subjects can be compared and analyzed. We have developed a mapping method for this purpose. Before applying this mapping method, all local evaluation metrics produced for each subject described above were mapped onto a single surface first.

2.E.1. 3D local evaluation metrics map

All of the local evaluation metrics were mapped to the mean surface generated from manual segmentation, which we refer as 3D local metrics map hereafter. At the end of the mapping operations, each point on the 3D local metrics map of subject j was equipped with the following list of local metrics:

ΔS_j: local mean distance between mean manual and algorithm segmentation [Eq. (1)];

$T_{\Delta S_j}$: results of point-by-point T-tests for ΔS [Eq. (3)];

$\mathrm{SD}_{j,A}$: local standard deviation for segmentations generated by algorithm;

$\mathrm{SD}_{j,M}$: local standard deviation for segmentations generated manually;

$\mathrm{SE}_{\Delta S_j}$: standard error of ΔS_j [Eq. (2)]; and

$\nu_{\Delta S_j}$: degrees of freedom of $T_{\Delta S_j}$ [Eq. (4)].

2.E.2. Generation of 3D standard local metrics map

A physician selected a template surface that has a more regular shape with average volume of the ten subjects in this study. The 3D subject-specific local evaluation metrics map described previously was mapped to the template surface. Iterative closest point[34] was first applied to rigidly align the 3D maps of different subjects with the template surface. The coherent point drift (CPD) described previously[35] was subsequently performed for nonrigid alignment. An example of the

FIG. 5. An example showing the CPD alignment operation. (a) and (b) show the ventricular surface of a subject and the template surface, respectively. The surface displayed in (a) was aligned with the template surface shown in (b) using CPD, resulting in the surface represented by the black mesh in (c), which are displayed together with the template surface (solid surface) for comparison.

CPD alignment operation is shown in Fig. 5. Point-by-point correspondence between the aligned surfaces was established by the 3D symmetric correspondence described previously. The 3D local evaluation metrics map of each subject was mapped to the template surface according to this correspondence relationship. Figures 6(a) and 6(b) show two examples of the subject-specific 3D ΔS maps, which were mapped to the template surface and shown in Figs. 6(c) and 6(d).

2.E.3. Mean metric map for whole group of subjects

Each subject is now associated with a 3D standard local metrics map. At each point (p_i) on the standard map, the mean of a local metric M, denoted as $\overline{M}(p_i)$, was defined as

$$\overline{M}(p_i) = \frac{1}{N}\sum_{j=1}^{N} M_j(p_i), \tag{11}$$

where $M_j(p_i)$ represents the value of local metric M at point p_i of the 3D standard map generated for subject j and N is the total number of subjects involved in this study. The local metric M in this equation could be one of the two metrics: ΔS and $|\Delta S|$.

In particular, we denote the average of $\Delta S_j(p_i)$ over all ventricles (i.e., $j \in [1,10]$) by $\overline{\Delta S}(p_i)$, which is the average signed difference between algorithm and manual segmentations at point p_i for the whole group of subjects. To determine whether $\overline{\Delta S}(p_i)$ was significant different from 0, T-test was performed at each p_i with the T statistic and degrees of freedom given below,[36]

$$T_{\overline{\Delta S}}(p_i) = \frac{\overline{\Delta S}(p_i)}{\mathrm{SE}_{\overline{\Delta S}}(p_i)}$$

with

$$\mathrm{SE}_{\overline{\Delta S}}(p_i) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\mathrm{SE}_{\Delta S_j}(p_i)^2}, \tag{12}$$

$$\nu_{\overline{\Delta S}(p_i)} = \frac{\left(\sum_{j=1}^{N}\mathrm{SE}_{\Delta S_j}(p_i)^2\right)^2}{\sum_{j=1}^{N}\frac{\mathrm{SE}_{\Delta S_j}(p_i)^4}{\nu_{\Delta S_j}(p_i)}}, \tag{13}$$

where $\mathrm{SE}_{\Delta S_j}(p_i)$ and $\nu_{\Delta S_j}(p_i)$ have been defined in Sec. 2.E.1.

Each point in the 3D standard map was associated with two standard deviations. The pooled standard deviation was computed for each point on the 3D standard map. The pooled standard deviation of a local standard deviation metric, $\mathrm{SD}(p_i)$, was computed by

$$\mathrm{SD}(p_i) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\mathrm{SD}_j^2(p_i)}, \tag{14}$$

where $\mathrm{SD}_j(p_i)$ denotes the standard deviation metric at point p_i of the 3D standard map generated for subject j and N is the total number of subjects. Local pooled standard deviation was computed for SD_A and SD_M.

3. RESULTS

3.A. Summary of segmentation performance for all subjects

Table I shows the mean and standard deviation of ventricular volume measurement obtained for each subject based on repeated manual and algorithm segmentations. The ten

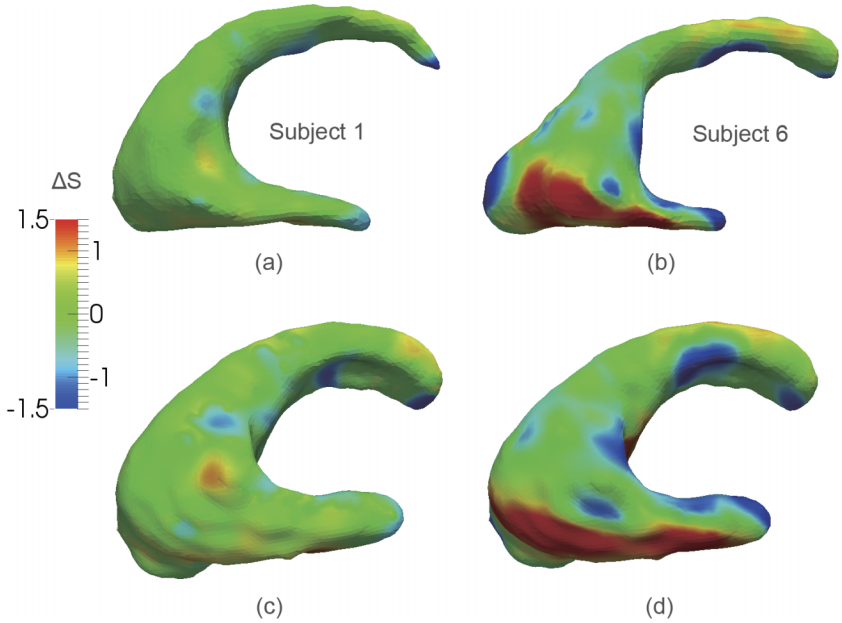

FIG. 6. Ventricular surfaces of (a) subject 1 and (b) subject 6 are color-coded and superimposed with ΔS. (Subjects were numbered according to Table II.) (c) and (d) show the template surfaces with ΔS color-coded and superimposed, respectively.

TABLE I. Mean and standard deviations (Std) of volume measurements in mm^3 obtained based on repeated manual and algorithm segmentations. DSC used to quantify the difference between the mean surfaces generated based on manual and algorithm segmentations was also tabulated.

	Manual		Algorithm		
Patient ID	Mean volume	Std	Mean volume	Std	DSC (%)
1	5 994	815	6 258	469	88.2
2	9 440	683	9 880	568	85.6
3	3 798	470	4 085	490	83.4
4	6 156	626	6 401	559	80.7
5	5 962	1010	6 567	702	76.4
6	5 263	792	6 079	371	74.8
7	10 000	2033	10 494	903	80.6
8	34 215	1910	29 739	1535	82.7
9	10 121	806	7 749	385	78.2
10	8 848	310	7 375	414	80.8

subjects involved in this study were identified by patient IDs ranging from 1 to 10, which are used hereafter to refer to a subject. The mean volumes were volumes of the mean surfaces estimated by STAPLE as described in Sec. 2.C. Dice similarity coefficients (DSCs) between the mean surfaces obtained by manual and algorithm segmentations were also tabulated. The average DSC for the ten subjects was 81.1(±4.1)%. As described in Sec. 2.D.1, ΔS, $|\Delta S|$, SD_M, and SD_A were computed on a point-by-point basis for each subject. Table II was constructed to summarize these local metrics for each subject. In this table, we list the average ΔS and $|\Delta S|$ as well as the pooled SD_M and SD_A computed over the whole ventricular surface for each subject.

3.B. Local metrics for a single subject mapped on the template surface

Subjects 1 and 6 were associated with higher DSC and lower DSC than average DSC, respectively. Here, we display and analyze the 3D local evaluation metric maps (Sec. 2.E) generated for these two subjects.

Figure 7 shows the ventricle surface for subject 1 with metrics introduced previously color-coded and superimposed.

TABLE II. Mean/pooled ΔS, $|\Delta S|$, SD_M, and SD_A in mm for the whole group of ten subjects. Mean ΔS and $|\Delta S|$ were computed as the mean of ΔS and $|\Delta S|$ over all the points on each subject, respectively. Pooled SD_M and SD_A were computed as the pooled standard deviation of manual and algorithm segmentation over all the points on each subject.

Patient ID	Mean ΔS	Mean $\lvert\Delta S\rvert$	Pooled SD_M	Pooled SD_A
1	0.02	0.43	1.38	0.95
2	0.02	0.81	1.25	0.79
3	0.04	0.75	0.94	0.68
4	0.02	0.54	0.86	0.73
5	0.01	0.69	1.80	0.87
6	0.14	1.14	0.92	0.50
7	0.02	1.23	2.75	1.62
8	−0.67	1.21	1.68	1.06
9	−1.25	1.45	1.34	0.61
10	−0.61	0.84	1.17	0.74

The mesh displayed in Fig. 7(a) represents the mean surface generated from six manual segmentations and the solid surface represents the mean surface of algorithm segmentations. Figure 7(b) shows the signed ΔS mapped on the manually segmented mean surface. This map shows that the algorithm oversegmented the ventricle under the inferior horn as pointed to by the solid arrow shown in Fig. 7(b) and undersegmented the tip of the anterior horn as pointed to by the dotted arrow shown in Fig. 7(b). Figures 7(c) and 7(d) show the local segmentation standard deviation associated with algorithm (SD_A) and manual methods (SD_M), respectively, superimposed on the manually segmented mean surface. Visual comparison of these two figures shows that the local segmentation standard deviation for manual segmentation is higher than that for algorithm segmentation. Both figures show elevated standard deviations at the inferior horn and atrium of the lateral ventricle. Figure 8 shows the 3D ultrasound images from which ventricular boundaries displayed in Fig. 7 were segmented. This figure shows that the boundary of the inferior horn is not well defined in the ultrasound image, which explains the elevated segmentation variability at this region for both manual and algorithm segmentations. Figure 7(e) shows the results of the point-by-point T-test for ΔS, in which white indicates statistical significant difference and the black indicates otherwise. Figure 9 shows the template surfaces with ΔS, SD_A, and SD_M and T-test superimposed. The surface area of subject 1 is 1888.3 mm^2. The ROI-based statistical framework described in Sec. 2.D.2 was applied on the whole ventricle surface of this subject, giving a probability distribution of

$$P\left(\sup_{p\in\Omega} T(p) > 8.63\right) = P\left(\sup_{p\in\Omega} T(p) < -8.63\right) \approx 0.025. \quad (15)$$

The maximum and minimum T values obtained for this subject were 10.97 and −7.89, indicating a significant difference between the manual and algorithm segmentations of this subject at a significant level of 5%.

Figure 10 shows Subject 6 with local metric color-coded and superimposed. As shown in Fig. 10(b), the algorithm over-segmented the region extending from the inferior to the posterior horn [red region in Fig. 10(b)] and under-segmented the tip of the inferior horn. Figure 11 shows the 2D contours in three views with the over-segmented location pointed by a solid arrow and under-segmented location pointed to by a dotted arrow. Figure 10(e) shows the results of T-tests for ΔS. In addition to the over-segmented region discussed above, Fig. 10(e) also shows significant local difference in the body of the ventricle extending from the anterior horn (gray arrows). Local segmentation variability of both manual and algorithm in this region was lower and a smaller local difference was enough to be detected as statistically significant. The local accuracy and variability metrics were mapped on the template surface and shown in Fig. 12. The surface area of Subject 6 is 1574.6 mm^2. The ROI-based statistical framework was applied on the whole ventricle surface of this subject, giving a probability distribution of

$$P\left(\sup_{p\in\Omega} T(p) > 9.81\right) = P\left(\sup_{p\in\Omega} T(p) < -9.81\right) \approx 0.025. \quad (16)$$

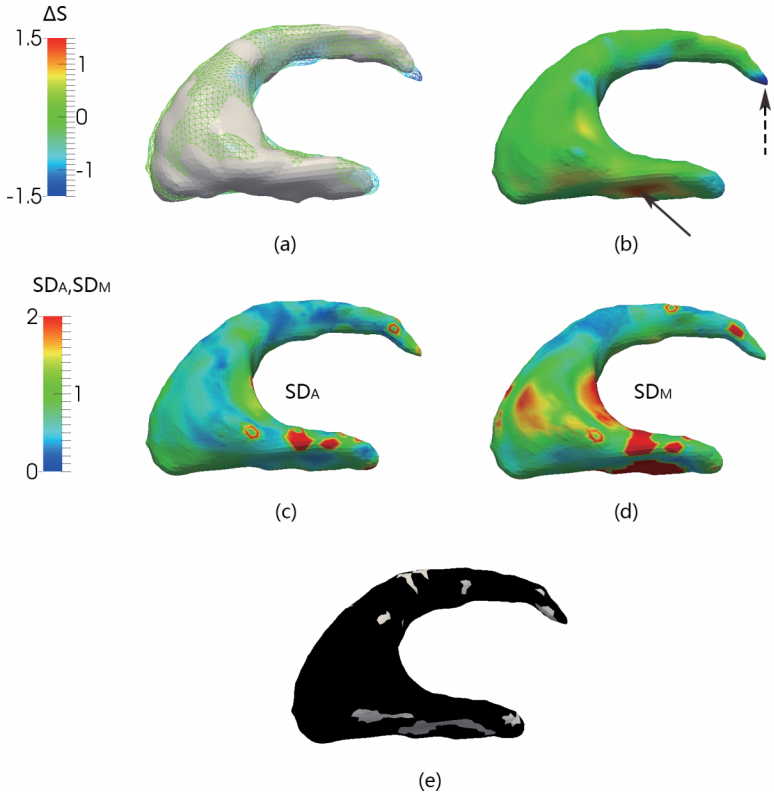

FIG. 7. Segmentation variability evaluation for subject 1. (a) shows the mean surface of algorithm segmentation and the mesh represents the manual segmentation surface with ΔS color-coded and superimposed. (b) ΔS, (c) SD_A, (d) SD_M, and (e) statistical significance (white indicates statistically significance, while black indicates otherwise) are color-coded and superimposed on the manual surface.

The maximum and minimum T values obtained for this subject were 19.63 and −17.23, indicating a significant difference between the manual and algorithm segmentations of this subject at a significance level of 5%.

To demonstrate that the accuracy and variability of the algorithm can be improved with more accurate initialization, we added initial points at the inferior and interior horns, where segmentation accuracy is lower according to the metric

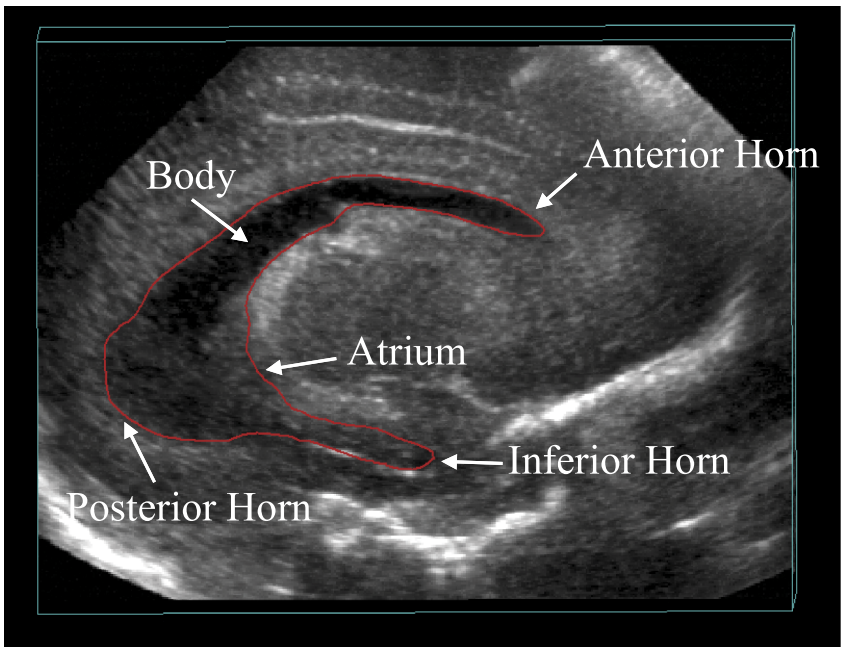

FIG. 8. A transverse image resliced from the 3D US image acquired for the ventricle displayed in Fig. 7. The red contour represents the mean manual boundary of the ventricle on the selected image slice.

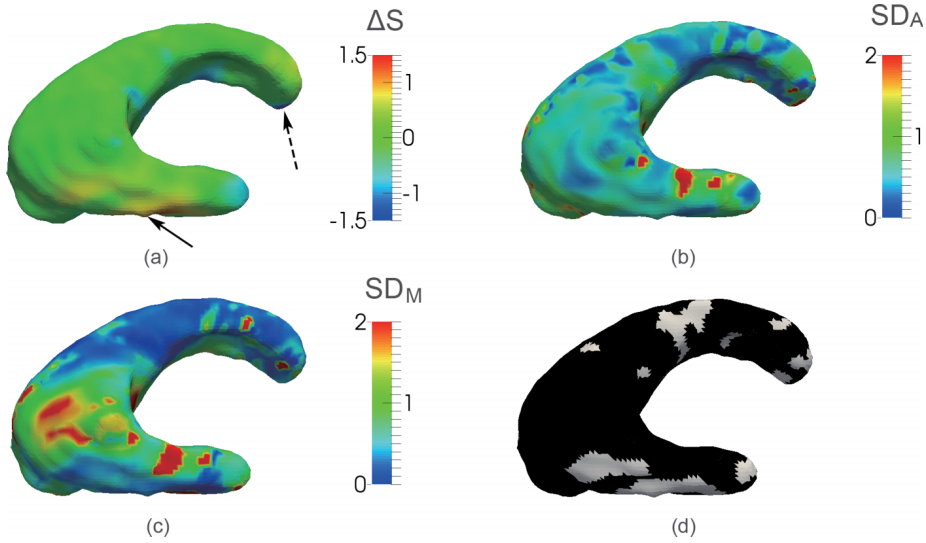

FIG. 9. (a) ΔS, (b) SD_A, (c) SD_M, and (d) T-test of subject in Fig. 7 are mapped to the template surface.

displayed in Fig. 10, to initialize the segmentation algorithm. Figure 13 shows the local error and variability of the algorithm segmentation after the editing of the initialization labels. The DSC increased from 74.76% to 87.15% while the mean $|\Delta S|$ decreased from 1.14 mm to 0.30 mm. The ROI-based statistical framework was applied on the whole ventricle surface of this subject, giving a probability distribution of

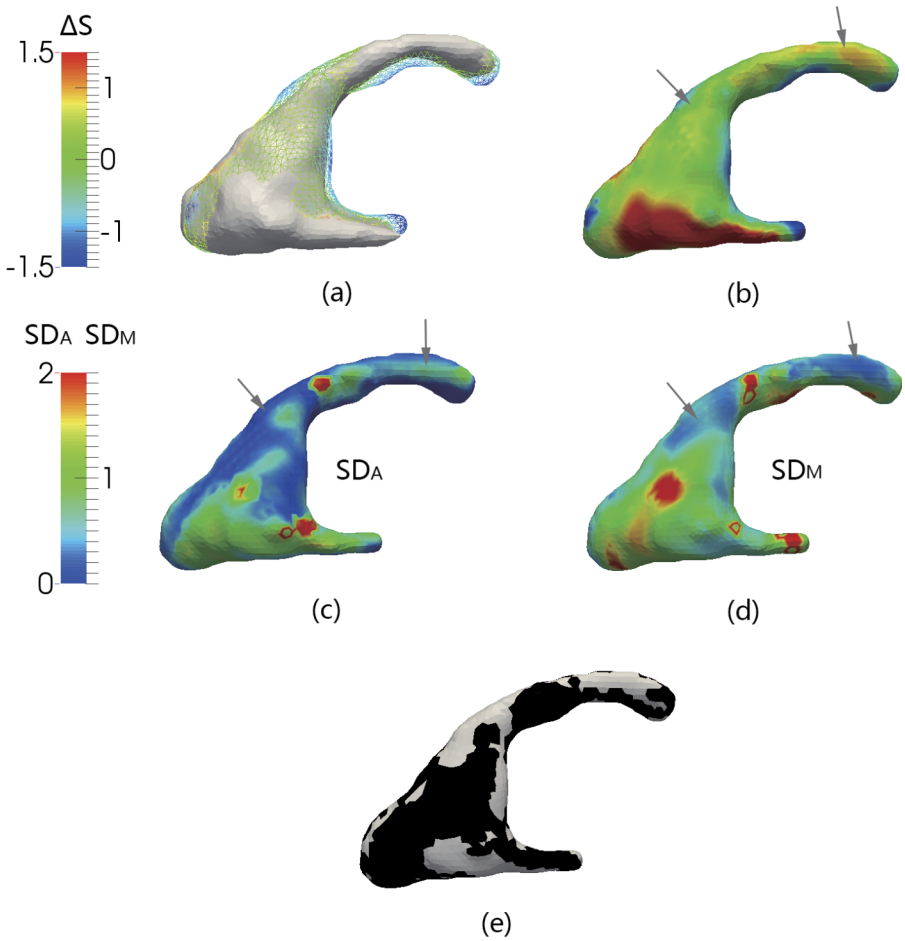

FIG. 10. Segmentation variability evaluation for subject 6. (a) shows the mean surface of algorithm segmentation and the mesh represents the manual segmentation surface with ΔS color-coded and superimposed. (b) ΔS, (c) SD_A, (d) SD_M, and (e) statistical significance (white indicates statistically significance, while black indicates otherwise) are color-coded and superimposed on the manual surface.

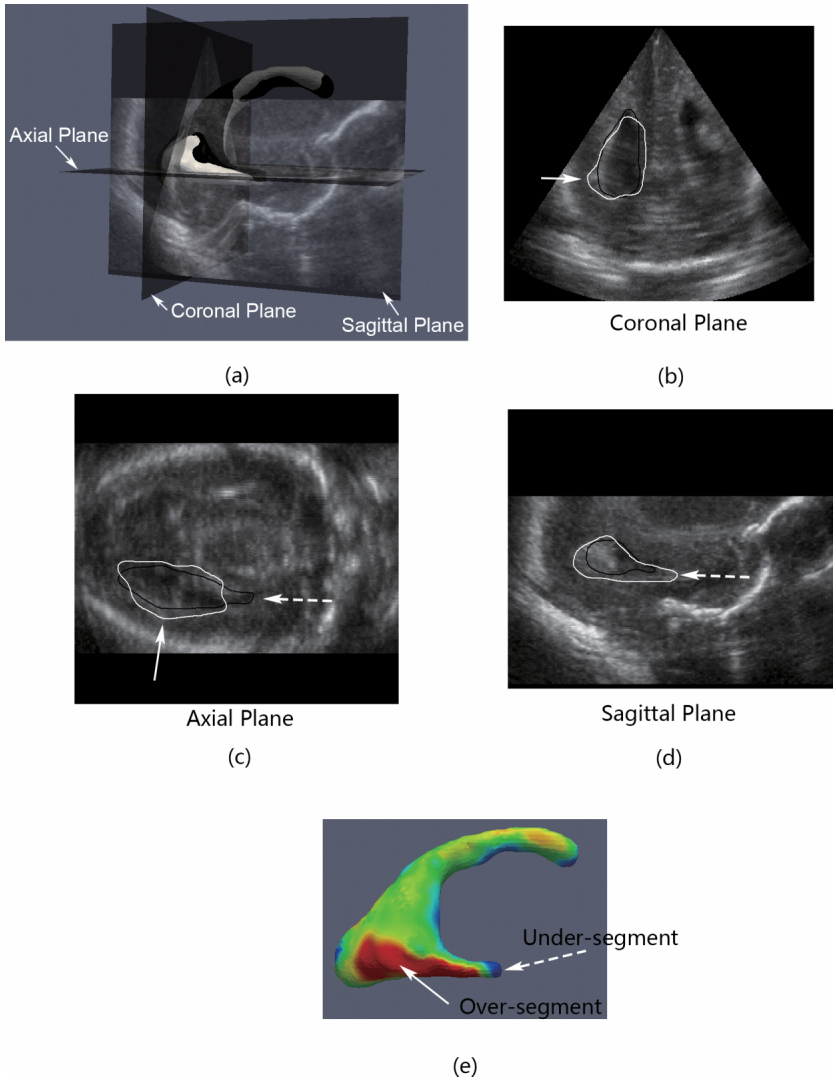

FIG. 11. (a) shows manual (black) and algorithm (white) based segmented surfaces in the 3D US image for the subject described in Fig. 10. Two dimensional image slices and the segmentation contours (black: manual and white: algorithm) on (b) coronal, (c) axial, and (d) sagittal plane are displayed. (e) shows the manual surface with ΔS color-coded and superimposed. The solid arrows point to the areas with oversegmentation while dotted arrows indicate undersegmentation.

$$P\left(\sup_{p\in\Omega} T(p) > 11.52\right) = P\left(\sup_{p\in\Omega} T(p) < -11.52\right) \approx 0.025. \tag{17}$$

The maximum and minimum T values were 12.40 and -13.20, still indicating that there is a significant difference between manual and algorithm segmentations, but the magnitudes of T values were reduced.

3.C. Regional analysis of segmentation errors

The anterior, inferior, and posterior horns were extracted from the ventricle surface of subject 1 as shown in Fig. 14. Regional statistical analyses were performed independently in the three subregions. Table III listed the thresholds, maximum, and minimum T values for them, respectively. For both anterior and inferior horns of this subject, regional analysis indicated significant ΔS between algorithm and manual segmentations while posterior horn showed no significant change.

3.D. Results on whole group of subjects mapped on the template surface

Figures 15 and 16 show the template surface with average ΔS, $|\Delta S|$, pooled standard deviations of algorithm (SD_A), and manual methods (SD_M) superimposed. Figure 17 shows the result of T-test performed on each point on the template according to Eqs. (12) and (13).

Figure 15(a) shows that the algorithm undersegmented regions in the neighborhood of the tips of the anterior, inferior, and posterior horns. High local segmentation variability is observed for both manual and algorithm segmentations in the atrium extending from the inferior to the posterior horn. This high variability was caused by the poor boundary definition in the neighborhood of the posterior and inferior horns in

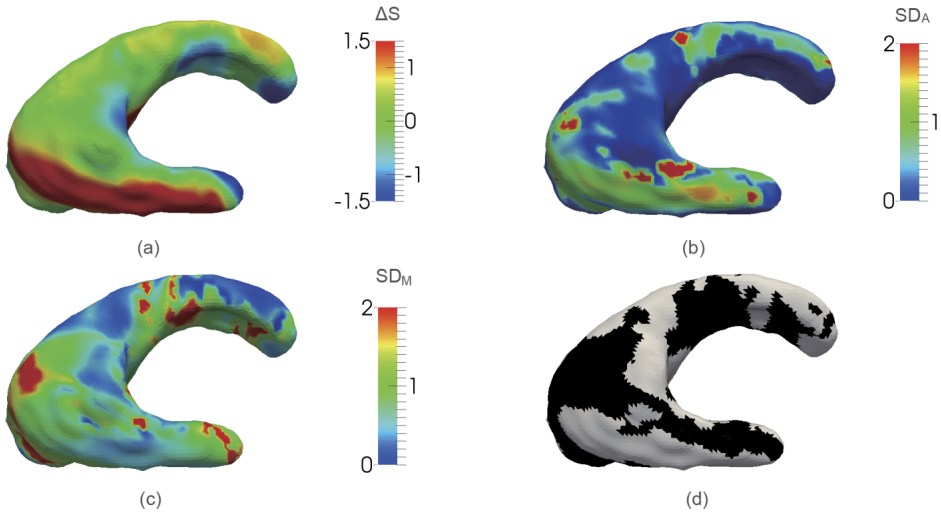

FIG. 12. (a) ΔS, (b) SD_A, (c) SD_M, and (d) T-test of subject in Fig. 10 are mapped to the template surface.

the 3D ultrasound image as shown in Fig. 8. The segmentation variability associated with manual segmentation is higher than algorithm segmentation over the whole ventricle, which was consistent with the volumetric evaluation performed in Qiu *et al.*[9] It is also not unexpected that $\overline{\Delta S}$ was not statistically significant at the inferior and posterior horns due to high segmentation variability at these regions. The ROI-based analysis introduced in Sec. 2.D.2 was applied for the group level statistical analysis on left and right lateral ventricles. The threshold, minimum, and maximum T values were listed as in Table IV, which indicated that there were significant differences between algorithm and manual segmentations for right ventricle as the minimum T value exceeded the threshold but not for the left lateral ventricle.

3.E. Result of group-level statistical analysis versus subject number

To determine whether ten subjects were sufficient in this statistical segmentation evaluation study, a good strategy would be to perform the group-level analysis (Sec. 2.E.3) using different number of subjects and evaluate whether the statistical results converge as the sample size approaches ten. Here, we performed the group-level pointwise analysis using seven to ten subjects [Eq. (12)] for both the left and right ventricles. When the subject number, N, was less than 10 (i.e., $7 \le N \le 9$), we performed the analyses for all C_{10}^{N} combinations of subjects. In each combination of subjects, we computed the percentage of points with $\overline{\Delta S}$ significantly different from 0. Table V reports the average of this percentage over all combinations of subjects. The results indicated that the average percentage converged as the sample size approached ten.

4. DISCUSSION

This paper focused on developing a statistical framework to answer the question of whether the neonatal ventricles segmented from 3D ultrasound images using an algorithm

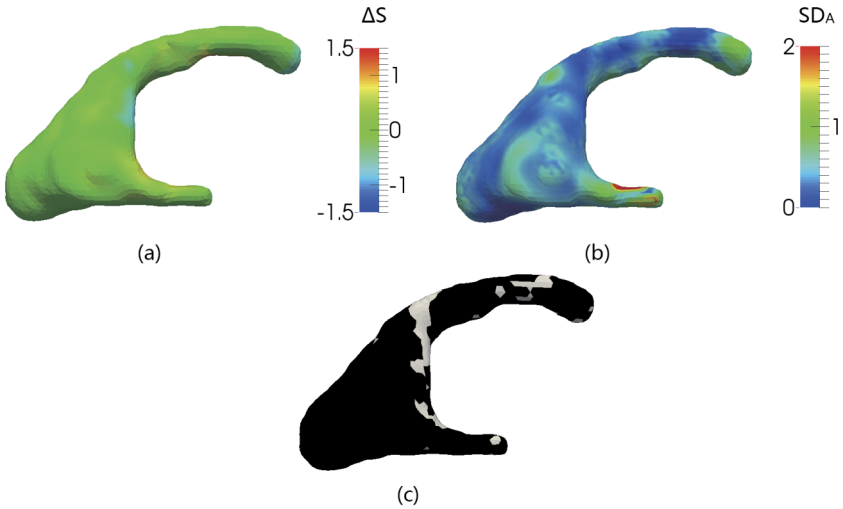

FIG. 13. Segmentation accuracy and variability with new initialization strategy according to the results obtained in Fig. 10. (a) shows the local signed distance between algorithm and manual segmentation, (b) shows the standard deviation of algorithm, and (c) shows local statistical significance (white indicates statistically significance while black indicates otherwise).

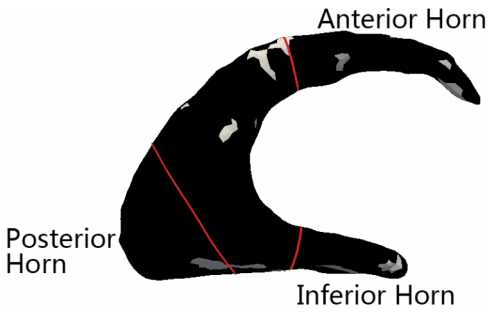

FIG. 14. Three horns were manually extracted from the 3D ventricle surface of subject 1 to demonstrate the regional statistical analysis. The red lines represent the boundaries of the three horns.

were significantly different from manual segmentation, which serves as the surrogate ground truth in segmentation evaluation. The paper introduces statistical tests to answer this question on the subject and the group levels.

In the subject-level analysis, the framework successfully addressed the challenges on (a) the requirement of a surface correspondence algorithm used to match the algorithm and manual segmentations on a point-by-point basis and (b) the estimation of segmentation variability from repeated segmentation. In particular, we addressed limitations of a previously described symmetric correspondence algorithm[30,37] which produces error when surfaces to be matched intersect each other. We developed a multiresolution approach for surface matching in which we first divided the surfaces into patches

TABLE III. Threshold, maximum, and minimum T values for three horns extracted from the ventricle surface as shown in Fig. 14.

Regions	Threshold T	Maximum T	Minimum T
Anterior horn	6.15	6.28	−5.26
Inferior horn	6.32	10.97	−4.46
Posterior horn	6.51	4.13	−3.18

based on intersecting lines. The patchwise matching operation was followed by a pointwise matching operation between each pair of patches. The mismatch problem associated with the previously described symmetric correspondence algorithm was avoided, because pointwise matching was only applied on nonintersecting patches. With the point-by-point correspondence relationship established between the algorithm and manual segmentations, pointwise distance between the two segmented ventricles, the segmentation variability in algorithm, and manual segmentations were obtained, from which a pointwise T-statistic can be derived to determine whether the algorithm segmentation was significantly different from the manual segmentation on a point-by-point basis [Eqs. (2)–(4)].

As a comprehensive local accuracy and variability assessment of an algorithm should be based on the segmentation results for a group of subjects, we developed statistical analysis on the group level. The major challenge of this level of analysis was the requirement to adjust for the anatomic difference of ventricles in a population of subjects. In this paper, we applied

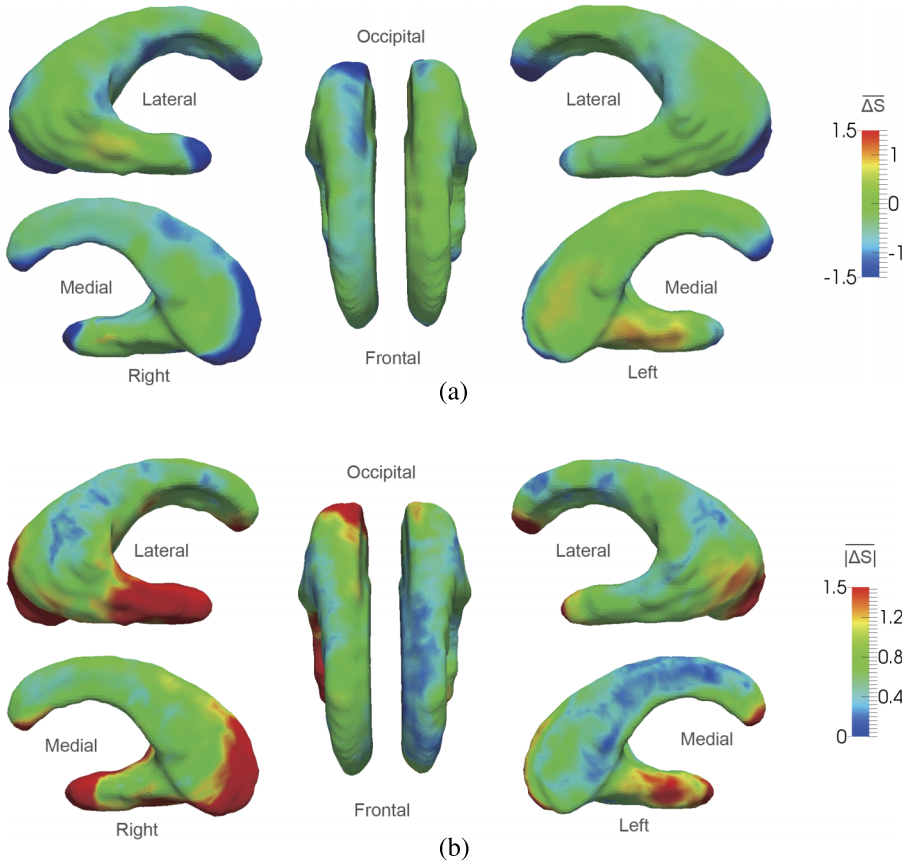

FIG. 15. Template surface with average (a) ΔS and (b) $|\Delta S|$ superimposed.

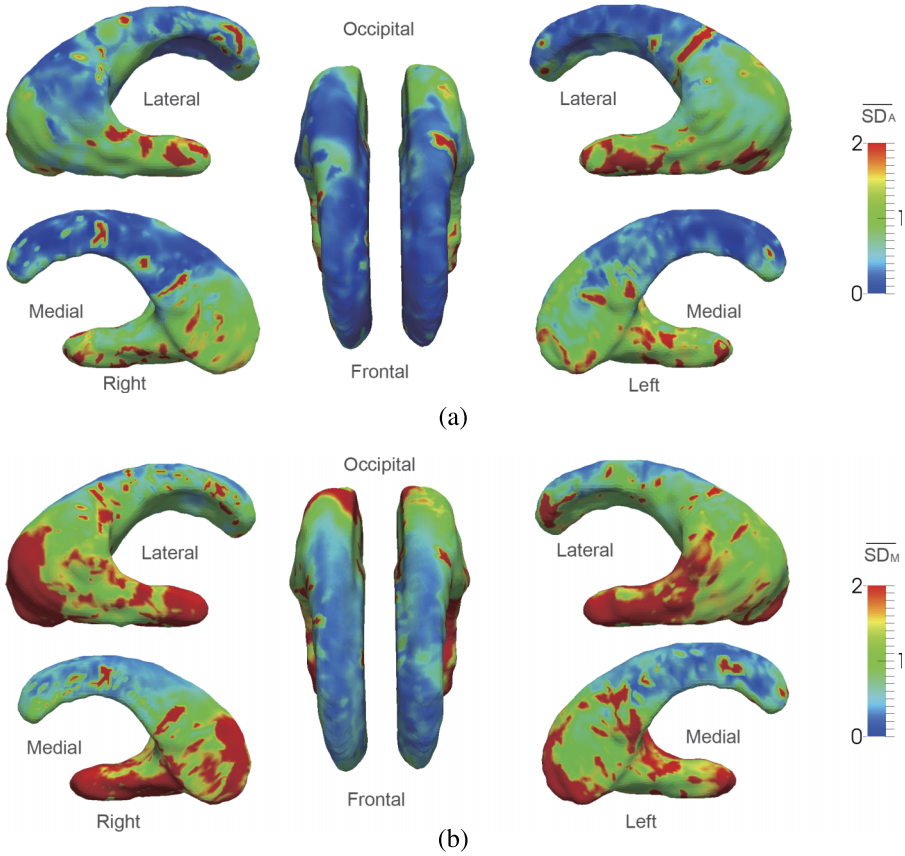

FIG. 16. Pooled standard deviations for (a) algorithmically (SD_A) and (b) manually (SD_M) segmentation are superimposed on template surface.

a nonrigid registration algorithm to align the ventricles of each subject with a template surface. On each point on the template surface, the subject-based accuracy and variability metrics can be averaged and pooled. Based on these metrics, we derived a T-statistic to determine whether the algorithm segmentations for the whole population of ventricles involved in this study were significantly different from the manual segmentations [Eqs. (12) and (13)].

An advantage of the proposed evaluation framework is that accuracy and variability metrics were computed on a point-by-point basis and can be superimposed on the ventricle surfaces. The segmentation accuracy and variability distribution maps allow for the identification of regions in which the algorithm segmentation is associated with low accuracy and high variability. For example, Fig. 10 shows that the segmentation algorithm greatly undersegmented the ventricles in the

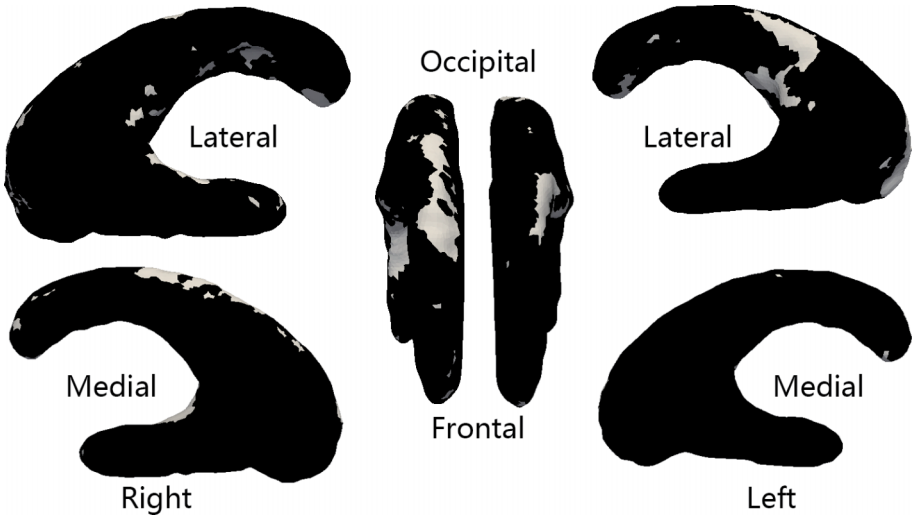

FIG. 17. Template surface with results of group-level T-test superimposed. White indicates statistically significance while black indicates otherwise.

TABLE IV. Threshold, maximum, and minimum T values for pooled statistics of left and right lateral ventricles.

Ventricle part	Threshold T	Maximum T	Minimum T
Left ventricle	7.75	3.86	−5.53
Right ventricle	6.58	1.25	−7.25

neighborhood of the posterior and inferior horns. To improve the performance of the algorithm, more initialization points can be introduced in these regions. Postediting mechanisms can also be introduced to highlight problematic regions based on the groupwise accuracy and variability maps. Users can then check and decide whether or not to edit the segmentation results in these regions. However, if only the subject-level accuracy and variability metrics were available in the framework, a user must carefully interpret the maps for the whole group of ten subjects in order to assess the overall performance of the algorithm. The amount of data available in these maps is too large (it would have required 50 subfigures in total, five subfigures for each subject as shown in Fig. 7) and may be too overwhelming for a user to assess the overall performance. The proposed framework provided a map displaying groupwise distribution map (Figs. 15–17) in order to summarize the accuracy and variability metrics and therefore facilitate interpretation of the algorithm in segmenting the whole population of subjects.

In addition to making a point-by-point statistical conclusion, conclusions regarding a subregion of the ventricle or even the whole ventricles can be made. The ΔS distributions can be modeled as Gaussian random fields with FWHM quantifying the spatial correlations between neighboring vertices on the ventricle surface.[32,38] Statistical distribution for the supremum of the T-statistic over a subregion or the whole ventricle was well-established (Table II in Worsley *et al.*[39]) that would allow statistical conclusions on the difference between algorithm and manual segmentations to be drawn on a regional and global scales both on the subject and group levels. The statistical conclusions obtained at different scales should be interpreted as a whole in order to develop an understanding of segmentation accuracy in different resolutions.

Although the proposed framework is developed for evaluating local accuracy and variability in segmenting the lateral ventricle from the 3D US images, it could be applied to assess the longitudinal local changes of lateral ventricles. Notably, deformation-based morphometry approaches have been proposed to monitor longitudinal changes.[12,13,16] However, the applications of these pipelines involved MRI brain images and required longitudinal changes to be small.[17] Monitoring longitudinal changes in neonatal ventricles with IVH pose unique challenges as the ventricle grows rapidly within a few days, which can be addressed by the surface-based correspondence algorithm we introduce in this paper. For example, the volume of neonatal ventricle in Fig. 18 increased from 6.84 to 10.61 cm^3 in three days. In a longitudinal study, the ventricle would be imaged at baseline and a follow-up session. After the surfaces segmented from the two ultrasound scans are registered, the framework described in this paper can be used to compute longitudinal local surface change. Figure 18(a) shows two left lateral ventricle surfaces of a patient at baseline and follow-up (three days later). The patient was diagnosed with bilateral grade III IVH and ventricle dilation was expected. This figure also shows that the symmetric correspondence algorithm introduced in this paper was able to match the two surfaces with large local difference [see inset plot in Fig. 18(a)]. Figure 18(b) shows the local surface change of the lateral ventricles for this subject. We observed that the most pronounced dilation occurred in the posterior horn region. There have been evidences showing connections between neuropsychological consequences and ventricular volume.[2,40] Patients with larger posterior horn demonstrated poorer visuospatial ability and it may be due to deformation or other hydrocephalus-related change in the posterior cortex.[2] Poor performance on motor functions has also been reported. It has been hypothesized that hydrocephalus impairs motor functions by deforming the cerebellum and stretching the corpus callosum.[2,41] However, to the best of our knowledge, only the lateral ventricle size has been quantified[42,43] and correlated with the long-term outcome of hydrocephalus, whereas the location where ventricular dilation occurs has never been considered. The proposed framework could be applied to quantify local ventricular dilation and may provide important data to improve the current understanding on the different effects of dilation occurring in different regions.

This statistical framework could also be applied to evaluate segmentation performance evaluation and monitor longitudinal changes of organs that can be described by a closed surface, such as corpus callosum, lung, and prostate. Also, our framework could be used as a metric for evaluating the performance of an observer in training for ventricle segmentation from 3D US images. In our group, each observer was trained to delineate contours under the supervision of physicians. The current practice is that the segmentation generated by the observer in training was validated against those segmented by an experienced physician based on intraclass correlation (ICC) of volume measurements. However, the volume difference does not reflect regional difference between segmentations performed by the observer and physicians. Using our proposed framework, the differences between segmentation performed by an observer in training and an experienced observer can be visualized directly on the 3D map, which gives detailed feedback for the observer in training to improve his/her performance.

TABLE V. Average percentage of points with $\overline{\Delta S}$ significantly different from 0 when different number of subjects were involved in the group-level statistical analysis.

	Average percentage of points with $\overline{\Delta S}$ significantly different from 0	
Number of selected subjects	Left ventricle (%)	Right ventricle (%)
7	9.7	12.4
8	9.5	11.8
9	9.6	12.2
10	9.2	11.9

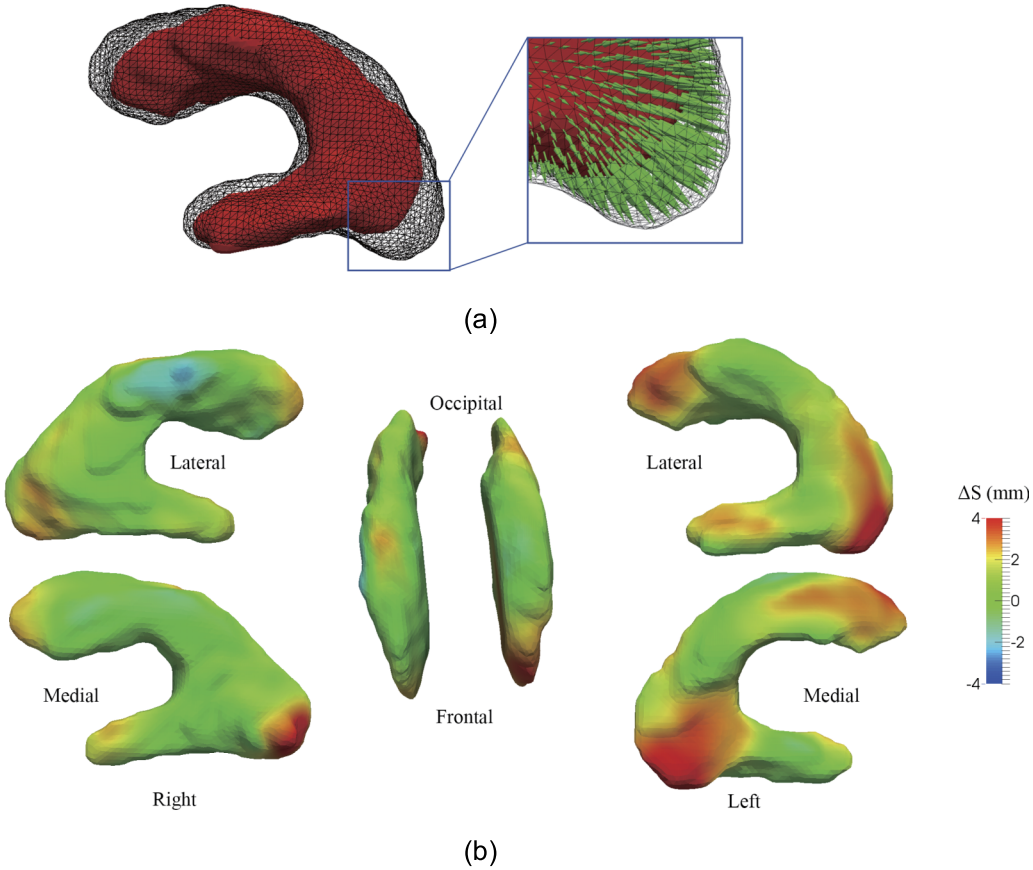

FIG. 18. (a) shows rigid registration of left lateral ventricle surfaces of a subject with grade III IVH obtained at baseline (solid surface) and follow-up (black mesh). The arrows connect pairs of corresponding points. (b) shows the 3D surfaces with local surface change color-coded and superimposed.

One limitation of this framework is that we used the line connecting the corresponding points on mean surfaces to intersect with six algorithms and six manual surfaces in order to establish the pointwise segmentation variability, but there is no guarantee that the line would intersect all segmented surfaces. In such situation, the surface without intersection would not be involved in the local standard deviation computation. In our studies, this situation was very rare. The line intersects less than six algorithm surfaces at 36 points and less than six manual surfaces at 57 points out of 25 082 points involved in this study (approximately 2500 points × 10 subjects).

Another limitation is that although we have validated the template surface mapping technique visually as demonstrated in Figs. 9 and 12, we have not performed a quantitative error analysis of the surface registration technique. Quantitative error evaluation involves introducing landmarks in the 3D US image. There has been no consensus on criteria for identifying landmarks on 3D ventricular ultrasound images. The irregular shape of neonatal ventricles caused by IVH makes the definition of landmarks an even bigger challenge. Before such a standard of defining landmarks is available, we will have to rely on the smoothness and topological constraints imposed on the CPD algorithm. Moreover, the registration error may depend on the selection of the template surface. Although MRI brain templates are widely established,[44,45] standard template was not available for neonatal ventricle imaged by 3D ultrasound. Constructing such a template would require another full publication, which involves careful identification of landmarks by expert observers and quantitative evaluation of the registration framework.

Although registration error may be introduced when morphing individual ventricles to the template surface, the effect of registration error on the statistical results generated in the group-level analysis [Eq. (12)] can be reduced by smoothing ΔS on each ventricle surface before mapping to the template surface. Suppose the point x in ventricle surface j should be mapped to p_i on the template surface, but due to registration error, the point y was mapped to p_i instead. Error introduced to the numerator of $T_{\overline{\Delta S}(p_i)}$ is a function of the difference between $\Delta S_j(x)$ and $\Delta S_j(y)$, which would be reduced upon smoothing. Smoothing will also reduce $\mathrm{SE}_{\Delta S_j}$ by a factor related to the FWHM (Appendix A in Forman *et al.*[38]). Hence, the error introduced to the denominator of $T_{\overline{\Delta S}(p_i)}$, which is a function of the difference between $\mathrm{SE}_{\Delta S_j}(x)$ and $\mathrm{SE}_{\Delta S_j}(y)$, will also be reduced. Although smoothing reduces the effect of registration error in the group-level statistical analysis, it would blur the fine details of the $\overline{\Delta S}$ distribution pattern. Hence, the FWHM chosen should be large enough to reduce the effect of registration error of $T_{\overline{\Delta S}(p_i)}$ to a predefined low level, yet not too large so that local features of the $\overline{\Delta S}$ distribution can still be identified. A future study would be required to quantify the amount of registration error and determine an appropriate

FWHM to reduce the effect of this registration error while preserving local information in the $\overline{\Delta S}$ distribution.

5. CONCLUSION

We developed a framework for evaluating, visualizing, and summarizing the local accuracy and variability of a segmentation algorithm on the subject and group levels. First, we developed a workflow to evaluate pointwise distances between manual and algorithm segmentations, and local variability of each of the segmentation methods based on repeated segmentations for each subject. This stage of analysis was carried out for each subject and thus was referred to as the subject-level analysis, which involves (a) reconstructing surfaces based on manual segmentations on sagittal slices, (b) computing a mean surface for each of the manual and algorithm segmentation methods based on repeated segmentations, (c) establishing pointwise symmetric correspondence between the manual and algorithm mean surfaces, (d) computing pointwise distance between the two surfaces, as well as segmentation variability of each segmentation method, and (e) performing T-test for each pair of corresponding point to establish whether or not the pointwise distance is statistically significant. The result can be superimposed on either the manual or algorithm mean surface for visualization and interpretation. This local accuracy and variability can be used to improve algorithm performance. To evaluate the segmentation performance on a whole group of ventricles, a nonrigid point set registration algorithm was applied to morph the ventricular surface of each subject to a ventricle template. We developed a statistical test based on segmentation accuracy and variability averaged over the whole group of subjects to determine, at each point of the template surface, whether the algorithm segmentations for the whole population of ventricles were significantly different from the corresponding manual segmentations. A major advantage of the proposed evaluation is that the average segmentation accuracy, variability, and the result of the pointwise statistical analysis can be superimposed and displayed on the template surface to illustrate and summarize the performance of an algorithm segmentation method in a whole group of subjects and more importantly, to guide user interactions aiming to improve segmentation performance of an algorithm. Finally, in addition to pointwise statistical evaluation performed on the subject and group level, we performed statistical tests on the algorithm accuracy over subregions and the whole ventricle, in which spatial correlation of ΔS was taken into consideration. Statistical analysis over a ROI is useful for quantitative characterization of the algorithm performance in relation to the regions with artifacts in the 3D ultrasound image.

ACKNOWLEDGMENTS

Dr. Chiu is grateful for funding support from the Research Grant Council of the HKSAR, China (Project No. CityU 139713) and the National Natural Science Foundation of China (Grant No. 81201149).

[a]Y. Chen and W. Qiu contributed equally to this work.
[b]Author to whom correspondence should be addressed. Electronic mail: bcychiu@cityu.edu.hk

[1]A. R. Synnes, L.-Y. Chien, A. Peliowski, R. Baboolal, and S. K. Lee, "Variations in intraventricular hemorrhage incidence rates among Canadian neonatal intensive care units," J. Pediatr. **138**, 525–531 (2001).
[2]M. Mataró, C. Junqué, M. A. Poca, and J. Sahuquillo, "Neuropsychological findings in congenital and acquired childhood hydrocephalus," Neuropsychol. Rev. **11**, 169–178 (2001).
[3]G. J. Harris, E. H. Rhew, T. Noga, and G. D. Pearlson, "User-friendly method for rapid brain and CSF volume calculation using transaxial MRI images," Psychiatry Res., Neuroimaging **40**, 61–68 (1991).
[4]M. I. Kohn, N. K. Tanna, G. Herman, S. Resnick, P. Mozley, R. Gur, A. Alavi, R. A. Zimmerman, and R. Gur, "Analysis of brain and cerebrospinal fluid volumes with MR imaging. Part I. Methods, reliability, and validation," Radiology **178**, 115–122 (1991).
[5]S. Horsch, J. Bengtsson, A. Nordell, H. Lagercrantz, U. Ådén, and M. Blennow, "Lateral ventricular size in extremely premature infants: 3D MRI confirms 2D ultrasound measurements," Ultrasound Med. Biol. **35**, 360–366 (2009).
[6]L. M. Leijser, L. Srinivasan, M. A. Rutherford, S. J. Counsell, J. M. Allsop, and F. M. Cowan, "Structural linear measurements in the newborn brain: Accuracy of cranial ultrasound compared to MRI," Pediatr. Radiol. **37**, 640–648 (2007).
[7]J. H. Gilmore, G. Gerig, B. Specter, H. C. Charles, J. S. Wilber, B. S. Hertzberg, and M. A. Kliewer, "Infant cerebral ventricle volume: A comparison of 3-D ultrasound and magnetic resonance imaging," Ultrasound Med. Biol. **27**, 1143–1146 (2001).
[8]J. Kishimoto, S. de Ribaupierre, D. Lee, R. Mehta, K. St. Lawrence, and A. Fenster, "3D ultrasound system to investigate intraventricular hemorrhage in preterm neonates," Phys. Med. Biol. **58**, 7513–7526 (2013).
[9]W. Qiu, J. Yuan, J. Kishimoto, J. McLeod, Y. Chen, S. de Ribaupierre, and A. Fenster, "User-guided segmentation of preterm neonate ventricular system from 3-D ultrasound images using convex optimization," Ultrasound Med. Biol. **41**, 542–556 (2015).
[10]Y. Xia, Q. Hu, A. Aziz, and W. L. Nowinski, "A knowledge-driven algorithm for a rapid and automatic extraction of the human cerebral ventricular system from MR neuroimages," NeuroImage **21**, 269–282 (2004).
[11]P. Anbeek, K. L. Vincken, G. S. Van Bochove, M. J. Van Osch, and J. van der Grond, "Probabilistic segmentation of brain tissue in MR imaging," NeuroImage **27**, 795–804 (2005).
[12]C. Davatzikos, A. Genc, D. Xu, and S. M. Resnick, "Voxel-based morphometry using the RAVENS maps: Methods and validation using simulated longitudinal atrophy," NeuroImage **14**, 1361–1369 (2001).
[13]J. C. Lau, J. P. Lerch, J. G. Sled, R. M. Henkelman, A. C. Evans, and B. J. Bedell, "Longitudinal neuroanatomical changes determined by deformation-based morphometry in a mouse model of Alzheimer's disease," NeuroImage **42**, 19–27 (2008).
[14]J. Hodel *et al.*, "3D mapping of cerebrospinal fluid local volume changes in patients with hydrocephalus treated by surgery: Preliminary study," Eur. Radiol. **24**, 136–142 (2014).
[15]M. Chung, K. Worsley, T. Paus, C. Cherif, D. Collins, J. Giedd, J. Rapoport, and A. Evans, "A unified statistical approach to deformation-based morphometry," NeuroImage **14**, 495–606 (2001).
[16]J. P. Boardman, S. J. Counsell, D. Rueckert, O. Kapellou, K. K. Bhatia, P. Aljabar, J. Hajnal, J. M. Allsop, M. A. Rutherford, and A. D. Edwards, "Abnormal deep grey matter development following preterm birth detected using deformation-based morphometry," NeuroImage **32**, 70–78 (2006).
[17]M. K. Chung, K. J. Worsley, S. Robbins, T. Paus, J. Taylor, J. N. Giedd, J. L. Rapoport, and A. C. Evans, "Deformation-based surface morphometry applied to gray matter deformation," NeuroImage **18**, 198–213 (2003).
[18]F. Mao, J. Gill, D. Downey, and A. Fenster, "Segmentation of carotid artery in ultrasound images: Method development and evaluation technique," Med. Phys. **27**, 1961–1970 (2000).
[19]B. Chiu, E. Ukwatta, S. Shavakh, and A. Fenster, "Quantification and visualization of carotid segmentation accuracy and precision using a 2D standardized carotid map," Phys. Med. Biol. **58**, 3671–3703 (2013).
[20]J. D. Gill, H. M. Ladak, D. A. Steinman, and A. Fenster, "Accuracy and variability assessment of a semiautomatic technique for segmentation of the carotid arteries from three-dimensional ultrasound images," Med. Phys. **27**, 1333–1342 (2000).

[21]M. Styner, I. Oguz, S. Xu, C. Brechbühler, D. Pantazis, J. J. Levitt, M. E. Shenton, and G. Gerig, "Framework for the statistical shape analysis of brain structures using SPHARM-PDM," Insight J. **1071**, 242–250 (2006).
[22]Y. Wang, L. M. Lui, X. Gu, K. M. Hayashi, T. F. Chan, A. W. Toga, P. M. Thompson, and S.-T. Yau, "Brain surface conformal parameterization using Riemann surface structure," IEEE Trans. Med. Imaging **26**, 853–865 (2007).
[23]Y. Wang, J. Zhang, B. Gutman, T. F. Chan, J. T. Becker, H. J. Aizenstein, O. L. Lopez, R. J. Tamburo, A. W. Toga, and P. M. Thompson, "Multivariate tensor-based morphometry on surfaces: Application to mapping ventricular abnormalities in HIV/AIDS," NeuroImage **49**, 2141–2157 (2010).
[24]B. A. Gutman, X. Hua, P. Rajagopalan, Y.-Y. Chou, Y. Wang, I. Yanovsky, A. W. Toga, C. R. Jack, Jr., M. W. Weiner, and P. M. Thompson, "Maximizing power to track Alzheimer's disease and MCI progression by LDA-based weighting of longitudinal ventricular surface features," NeuroImage **70**, 386–401 (2013).
[25]A. Fenster, D. B. Downey, and H. N. Cardinal, "Three-dimensional ultrasound imaging," Phys. Med. Biol. **46**, R67–R99 (2001).
[26]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. Imaging **23**, 903–921 (2004).
[27]W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," ACM SIGGRAPH Comput. Graphics **21**, 163–169 (1987).
[28]G. J. Grevera and J. K. Udupa, "Shape-based interpolation of multidimensional grey-level images," IEEE Trans. Med. Imaging **15**, 881–892 (1996).
[29]A. G. Bors, L. Kechagias, and I. Pitas, "Binary morphological shape-based interpolation applied to 3-D tooth reconstruction," IEEE Trans. Med. Imaging **21**, 100–108 (2002).
[30]X. Papademetris, A. J. Sinusas, D. P. Dione, R. T. Constable, and J. S. Duncan, "Estimation of 3-D left ventricular deformation from medical images using biomechanical models," IEEE Trans. Med. Imaging **21**, 786–800 (2002).
[31]C. Quammen, C. Weigle, and R. M. Taylor, "Boolean operations on surfaces in VTK without external libraries," Insight J. **797**, 1–12 (2011).
[32]D. J. Hagler, A. P. Saygin, and M. I. Sereno, "Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data," NeuroImage **33**, 1093–1103 (2006).
[33]J. Cao and K. Worsley, "Applications of random fields in human brain mapping," in *Spatial Statistics: Methodological Aspects and Applications* (Springer, New York, NY, 2001), pp. 169–182.
[34]P. Besl and N. D. McKay, "A method for registration of 3-D shapes," IEEE Trans. Pattern Anal. Mach. Intell. **14**, 239–256 (1992).
[35]Y. Gao, J. Ma, J. Zhao, J. Tian, and D. Zhang, "A robust and outlier-adaptive method for non-rigid point registration," Pattern Anal. Appl. **17**, 379–388 (2014).
[36]F. E. Satterthwaite, "An approximate distribution of estimates of variance components," Biom. Bull. **2**, 110–114 (1946).
[37]B. Chiu, M. Egger, J. D. Spence, G. Parraga, and A. Fenster, "Quantification of carotid vessel wall and plaque thickness change using 3D ultrasound images," Med. Phys. **35**, 3691–3710 (2008).
[38]S. D. Forman, J. D. Cohen, M. Fitzgerald, W. F. Eddy, M. A. Mintun, and D. C. Noll, "Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold," Magn. Reson. Med. **33**, 636–647 (1995).
[39]K. J. Worsley *et al.*, "A unified statistical approach for determining significant signals in images of cerebral activation," Hum. Brain Mapp. **4**, 58–73 (1996).
[40]J. M. Fletcher, T. P. Bohan, M. E. Brandt, L. A. Kramer, B. L. Brookshire, K. Thorstad, K. C. Davidson, D. J. Francis, S. R. McCauley, and J. E. Baumgartner, "Morphometric evaluation of the hydrocephalic brain: Relationships with cognitive development," Child. Nerv. Syst. **12**, 192–199 (1996).
[41]M. Dennis, C. R. Fitz, C. T. Netley, J. Sugar, D. C. Harwood-Nash, E. B. Hendrick, H. J. Hoffman, and R. P. Humphreys, "The intelligence of hydrocephalic children," Arch. Neurol. **38**, 607–615 (1981).
[42]S. Jary, A. De Carli, L. A. Ramenghi, and A. Whitelaw, "Impaired brain growth and neurodevelopment in preterm infants with posthaemorrhagic ventricular dilatation," Acta Paediatr. **101**, 743–748 (2012).
[43]L. M. Fox, P. Choo, S. R. Rogerson, A. J. Spittle, P. J. Anderson, L. Doyle, and J. L. Cheong, "The relationship between ventricular size at 1 month and outcome at 2 years in infants less than 30 weeks gestation," Arch. Dis. Child. Fetal Neonatal Ed. **99**, F209–F214 (2014).
[44]K. Kazemi, H. A. Moghaddam, R. Grebe, C. Gondry-Jouet, and F. Wallois, "A neonatal atlas template for spatial normalization of whole-brain magnetic resonance images of newborns: Preliminary results," NeuroImage **37**, 463–473 (2007).
[45]F. Lalys, C. Haegelen, J. C. Ferre, O. El-Ganaoui, and P. Jannin, "Construction and assessment of a 3-T MRI brain template," NeuroImage **49**, 345–354 (2010).