How to Evaluate Medical Visualizations on the Example of 3D Aneurysm Surfaces

S. Glaßer[1], P. Saalfeld[1], P. Berg[2], N. Merten[1], B. Preim[1]

[1] Department for Simulation and Graphics, Otto-von-Guericke University Magdeburg, Germany
[2] Department of Fluid Dynamics and Technical Flows, Otto-von-Guericke University Magdeburg, Germany

Abstract

For the evaluation of medical visualizations, a ground truth is often missing. Therefore, the evaluation of medical visualizations is often restricted to qualitative comparisons w.r.t user preferences but neglects more objective measures such as accuracies or task completion times. In this work, we provide a pipeline with statistical tests for the evaluation of the user performance within an experimental setup. We demonstrate the adaption of the pipeline for the specific example of cerebral aneurysm surface visualization. Therefore, we developed three visualization techniques to compare the aneurysm volumes. Then, we present a single-factor, within-subject user study, which allows for the evaluation of these visualization techniques as well as the identification of the most suitable one. The evaluation includes a qualitative as well as a comprehensive quantitative analysis to determine statistically significant differences. As a result, a color-coded map surface view is identified as best suited to depict the aneurysm volume changes. The presentation of the different stages of the evaluation pipeline allows for an easy adaption to other application areas of medical visualization. As a result, we provide orientation to enrich qualitative evaluations by the presented quantitative analyses.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation – Line and curve generation G.3 Probability and Statistics Experimental Design

August 16, 2016

1 Introduction

Nowadays, many approaches for supporting the clinical expert as well as the clinical researcher regarding diagnosis or therapy involve a number of computer-supported segmentation, visualization and evaluation steps. In this paper, we explain how to evaluate comparative medical visualizations. Although many authors conduct a qualitative evaluation with medical domain experts, those results are hardly reproducible. Often, these domain experts are cooperation partners and co-authors of the presented work, where a subjective bias is hardly avoidable. Nevertheless, their medical knowledge is essential for specific application areas which may justify this procedure. Therefore, we do not neglect qualitative studies, but we want to include quantitative statistical analyses such that they can be easily adapted by other medical visualization researchers for a more comprehensive evaluation.

The segmentation of vessels with pathologic changes such as aneurysms or stenoses is an important research area. To create reproducible results as well as to reduce the work load of clinicians, automatic segmentations of vascular structures are desired. Due to patient-specific anatomies and pathologies, such automatic solutions remain challenging, and aiming for a general automatic segmentation framework is probably illusory [LABFL09]. Our application area is the visualization of cerebral aneurysms. Aneurysms bear the risk of rupture, which may cause severe consequences for the patients. For an improved intervention planning, patient-specific 3D surface models of the aneurysm and the surrounding vascular tree are extracted. They allow for the simulation of the internal blood flow [CCA+05, BRB+15] or the extraction of morphological parameters [LEBB09]. The results are included into the minimally invasive surgical plan as well as the post-processing applications within the clinical environment.

Our application scenario does not focus on the segmentation technique, but rather on the comparative visualization of different segmentation results. The employed surface meshes were extracted with a threshold-based segmentation, which can be successfully used for cerebral aneurysms [GBNP15]. However, during the segmentation process, the clinical expert requires feedback how parameters influence the segmentation results since small parameter changes may induce enormous changes of the surface mesh. To guide the clinical expert through the segmentation process, we developed three different comparative visualization techniques to show surface mesh variations.

Our quantitative and qualitative evaluation allows for the identification of the most suitable visualization technique. It comprises the visualization of five cerebral aneurysms, each approximated with three slightly different surface meshes. Our conducted user study determines which visualization technique is best suited to evaluate the perception of small changes in the aneurysm volumes. This is especially necessary when the clinical expert or medical researcher tunes the parameters of the segmentation process and requires feedback, whether the aneurysm extent increases or not. The presented concepts comprising the experimental setup, the study design, the study procedure as well as the statistical evaluation, can be easily generalized and thus, transferred to other medical visualization application areas. Our contributions are:

- We explain which statistical test is suitable for analysis of a user study and order them into a general pipeline for the qualitative and quantitative evaluation of medical visualizations.
- We use the application scenario of cerebral aneurysms to provide three techniques Vis_A, Vis_B, and Vis_C for the visualization of two similar but not identical aneurysm surface meshes, which mutually penetrate and overlap.
- Based on this example, we demonstrate how to employ the pipeline to determine which visualization technique is best suited for this application.

2 Related Work

In this section, we discuss related work for the qualitative and quantitative evaluation of visualizations with focus on the application area of aneurysms and vessels. We also refer to comparative visualizations of surface meshes extracted from medical image data.

In recent years, findings from psychophysical studies were incorporated to enhance 2D and 3D visualizations [BCFW08] influencing also the evaluation process of visualizations. For the assessment of a visualization's suitability and performance, user studies offer a scientifically sound method [KHI+03]. Isenberg et al. [IIC+13] provide a systematic review of the evaluation practices in visualization. They employ several evaluation categories and conclude that the *Qualitative Result Inspection* was most often used by all reviewed papers. Further emphasis on evaluation of algorithmic performance as well as an increasing trend in the evaluation for user experience and user performance were reported.

This finding is also reflected in medical visualizations. Often, a user study is carried out, where the participants provide a subjective rating of the novel algorithm. Gasteiger et al. [GNKP10] carried out a user study for their aneurysm visualization based on the participant's grade of satisfaction w.r.t. depth perception, spatial relationships, flow perception and surface shape. Subsequently, a more quantitative evaluation was presented by Baer et al. [BGCP11] for this visualization technique amongst two others. They conducted three controlled, task-based experiments and were able to determine statistically significant differences for the visualizations. Borkin et al. [BGP+11] also includes a formal quantitative user study to determine which visualization technique of the endothelial shear stress of coronary arteries is best suited. Hence, the experimental study provided by Díaz et al. [DRN+15] comprises a test setup to evaluate different shading techniques for volume data sets. Their evaluation included a quantitative statistical analysis as well. Also, perceptually motivated medical visualizations often include quantitative evaluations [PBC+16]. However, they focus on abstract information, e.g., depth perception, rather than comparing visualization techniques for a specific medical application area.

Visualizations of vessels are often depicted as 3D surfaces due to their complex and patient-individual shape [BFLC04, SOBP07, PO08]. Furthermore, overview visualizations are possible, e.g., the CoWRadar visualization for cerebral vessels [MMNG15]. Since we intend to employ aneurysm surface meshes for subsequent computational fluid dynamics (CFD) simulations and morphological analyses, we focus on 3D surface visualization methods. The depiction of cerebral aneurysms mostly involves the visual representation of hemodynamic parameters, e.g., scalar parameters are displayed via color-coded surface views [CSP10]. Gasteiger et al. [GNKP10] developed an illustrative visualization of aneurysms using a Fresnel shading to reveal the embedded blood flow. This work strongly motivated our visualization technique Vis_B.

Our comparative visualization is inspired by the image-based rendering of intersecting surfaces [BBF$^+$11]. This technique is based on the approach by Weigle and Taylor [WT05]. Next to the integration of additional local distance cues, they enabled interactive manipulation of the surfaces. Geurts et al. [GSK$^+$15] employed a visual comparison of medical segmentation results to allow for an evaluation of the segmentation quality. They provided additional information with landmark-based clustering to detect similar segmentation results. For the visualization itself, a color-coding of the surface was employed. There also exist illustrative approaches, e.g., the visualization presented by Carnecky et al. [CFM$^+$13]. However, we aim at a fast comparison of cerebral aneurysm volume. Therefore, we want to reduce the visual complexity and choose the concepts provided by Busking et al. [BBF$^+$11] as inspiration for our technique Vis_C.

Our visualization techniques show different segmentation results from the same patient which can be also interpreted as uncertainty visualization. Grigoryan and Rheingans [GR04] presented point-based probabilistic surfaces, which visualize surface models of medical structures such as tumors. Hence, the surface points are displaced to reflect the uncertainty at that point. The method by Pöthkow and Hege [PH11] comprises a feature-based visualization for isosurfaces with uncertainties. Their approach employs color-coding, glyphs and direct volume rendering. A taxonomy of uncertainty visualization approaches is provided by Potter et al. [PRJ12].

3 Medical Background and Image Data

Cerebral aneurysms are pathologic dilatations of the cerebral artery walls which may rupture and cause a subarachnoid hemorrhage with severe consequences for the patient. Treatment is carried out via endovascular intervention or neurosurgical clipping. However, the treatment itself may cause complications such as hemorrhages. The mortality rate associated with treatment is reported to be higher than the rupture rate of small asymptotic aneurysms [Wie03]. Thus, rupture risk assessment is an active clinical research area.

Rupture risk factors in clinical practice mainly comprise the aneurysm's morphology as well as the type of aneurysm, i.e., asymptomatic or symptomatic [WvdSAR07]. Hence, extraction of surface meshes for aneurysms provide additional information such as the evaluation of the ostium area (i.e., the orifice between the aneurysm sac and the parent artery) [LEBB09]. Further research directions involve the simulation of the internal blood flow since unstable and complex blood flow was correlated with increased rupture risk [CCA$^+$05, XNT$^+$11]. Again, a patient-specific surface mesh is the prerequisite for volume grid extraction and a subsequently CFD simulation.

For diagnosis of cerebral aneurysms, rotational angiography (RA) is considered as gold standard imaging method [GLR$^+$09] due to the high spatial resolution. Based on RA data, the 3D digital subtraction angiography (DSA) data sets are reconstructed. To obtain the slightly similar surface meshes, we exploit the reconstruction process of the RA data from the DSA suite (Siemens Artis zeego, Siemens Healthcare GmbH, Erlangen, Germany). Five patient-specific cerebral aneurysm data sets (D_1-D_5) were reconstructed using three different kernels: Hounsfield unit (HU) smooth, HU normal and HU sharp. The five aneurysms stem from five female patients with mean age of 49 years (range 45-59 years). One cerebral aneurysm was located at the anterior communicating artery, one at the posterior communicating artery, two at the segment of the internal carotid artery, and one at the bifurcation of the middle cerebral artery. Their size varied from 2.5 mm to 11.2 mm. All patients were treated with endovascular coiling.

4 Segmentation and Comparative Visualization of Cerebral Aneurysms

In this section, the aneurysm and ostium segmentation is explained. Afterwards, the three visualization techniques Vis_A, Vis_B and Vis_C are presented.

4.1 Segmentation of Aneurysm and Ostium

For each patient's RA data set, the three different reconstruction kernels yield three different DSA data sets. For each patient, a threshold-based segmentation was carried out for the HU normal reconstructed DSA image data. The resulting surface meshes are depicted in Figure 1. Next, the remaining reconstructions of the same patient were carried out such that they exhibit similar contours in a representative slice covering the aneurysm (see Fig. 2). Based on each threshold, the iso-surface is extracted and converted into the triangle surface mesh. Data inspection, threshold segmentation and mesh generation was carried out in MeVisLab 2.7 (MeVis Medical Solutions AG, Bremen, Germany). Hence, the segmentation was not the focus of our work and depending on the medical application, a fully automatic segmentation can be employed as well. For the purpose of our study, we required similar, but not identical aneurysm surface meshes, which could be successfully extracted with the threshold-based segmentation from different reconstructed RA data sets.

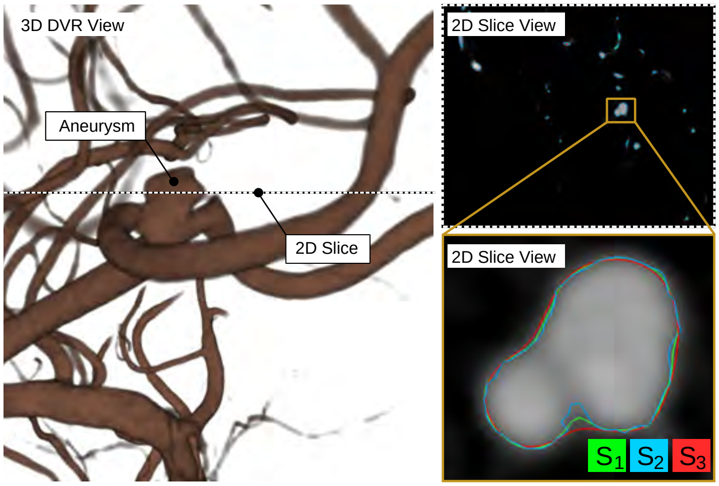
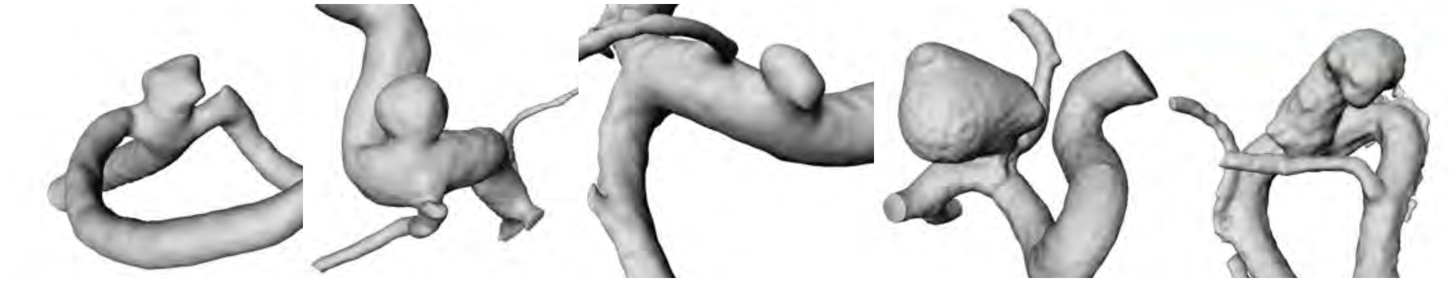


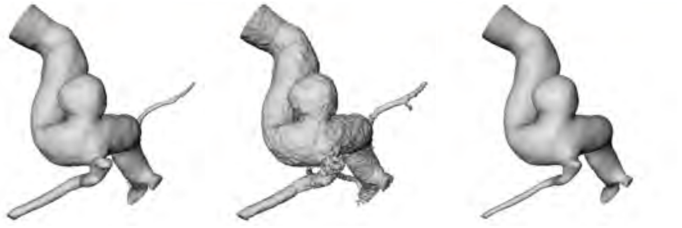
Figure 2: Segmentations of patient P_1. On the left, a direct volume rendering of the DSA data set is depicted. A 2D slice covering the aneurysm is shown on the right, its position is also highlighted in the 3D view. Thresholds for segmentations S_1-S_3 are selected such that similar segmentations are achieved, see bottom right. The resulting segmentation masks are color-coded.

Our visualizations focus on the comparison of the volume of each aneurysm without the surrounding vessel tree. Therefore, visual separation between aneurysm and parent vessel has to be provided. The ostium was manually extracted by defining a closed cutting line along the aneurysm surface mesh using Blender 2.74 (Blender Foundation, Amsterdam, The Netherlands). This cutting line was employed twice. First, we create a closed ostium surface by triangulating the cutting line. The aneurysm surface was cut with this



Patient P_1 Patient P_2 Patient P_3 Patient P_4 Patient P_5

Figure 1: Surface meshes of five patient data sets P_1-P_5 reconstructed with the HU normal kernel are shown.



P_2 S_1 P_2 S_2 P_2 S_3

Figure 3: For patient P_2, the three resulting segmentations S_1, S_2 and S_3 based on the different reconstruction kernels (HU normal, HU sharp, HU smooth) are depicted.

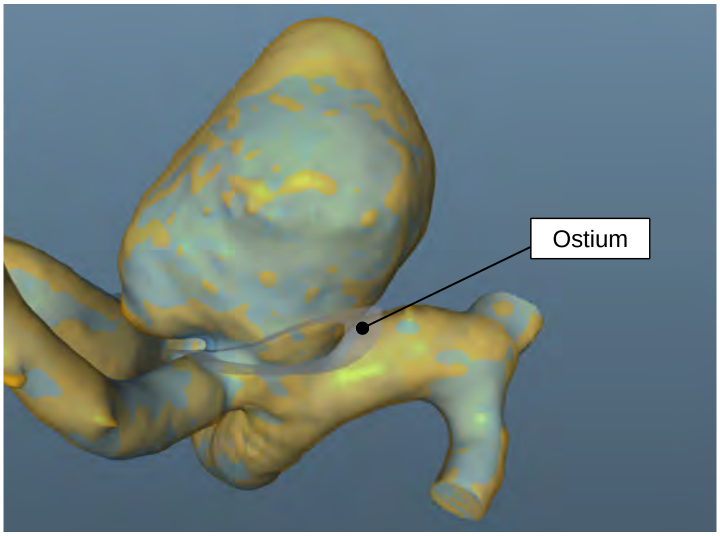


Figure 4: Depiction of the iso-surface view Vis_A. In case the surface mesh of A_{Ref} exceeds the surface mesh of A_{Comp}, the orange surface becomes visible. Otherwise, the cyan mesh is visible. The ruff-like structure provides information about the ostium.

surface to extract the aneurysm's volume for our evaluation. Second, the cutting line's vertices were extruded to create a ruff-like structure, which supports the participants of our user study. An automatic ostium segmentation was not the focus of this paper, but interested readers are referred to Neugebauer et al. [NLBP13].

4.2 Comparative Visualization Techniques

To evaluate differences of the aneurysm volume, we developed three visualization techniques: the iso-surface view Vis_A, the boundary-enhancing shading view Vis_B, and the color-coded map surface view Vis_C. Each technique shows two aneurysms, where the first one is referred to as A_{Ref}, i.e., the reference aneurysm, and the second one as A_{Comp}, i.e., the aneurysm for comparison. Note that the ordering of the aneurysms is important, and employing A_{Ref} first and A_{Comp} second yields a different visualization result than the usage of A_{Comp} first and A_{Ref} second. In the following, the visualization techniques will be described in more detail.

4.2.1 The Iso-Surface View - Vis_A

The iso-surface view is a rather straightforward direct visualization of the two surface meshes of the aneurysms A_{Ref} and A_{Comp}. It is realized in MeVisLab using the Open Inventor Library. For A_{Ref} an orange, and for A_{Comp} a cyan transparent surface mesh is simultaneously visualized with opacity values of 0.5 (see Fig. 4). Beyond mesh extraction, no further preprocessing is required.

4.2.2 The Boundary-Enhanced View - Vis_B

The second visualization technique Vis_B (see Fig. 5) is based on the Fresnel shading approach, which was successfully employed for aneurysm visualization comprising an inner blood flow visualization [GNKP10] or the outer vessel wall revealing the color-coded inner vessel wall [GLH+14]. This technique is also referred to as ghosted view or x-ray shading. Although we do not include additional information yet, e.g., the inner blood flow, we do integrate this visualization technique in our user study since we are interested in a possible extension of the visualization with the above-mentioned information in the future.

The opacity o for each surface mesh is assigned in the fragment shader and depends on the normal n and the viewing vector v:

$$o = 1 - (\vec{n} \cdot \vec{v})^f,$$

where f serves as edge-fall-off parameter. This parameter strongly influences the visualization of possible inner structures. We use an empirically determined value of $f = 0.7$. The visualization technique is realized in MeVisLab using the Open Inventor vertex and fragment shader modules, where the user can directly provide shader codes as input.

4.2.3 Map-Surface-View - Vis_C

In contrast to Vis_A and Vis_B, the map surface view visually provides quantitative information for the distance between A_{Ref} and A_{Comp}. For the gathering of the distance information, the estimation of the nearest vertex pairs from A_{Ref} and A_{Comp} is carried out. We calculate the normals of the A_{Ref} surface mesh and approximate the distance based on the intersection with A_{Comp}. The normals of A_{Ref} point inwards. If A_{Comp} is larger than A_{Ref}, the intersection in negative normal direction is nearer to A_{Ref}'s vertex than the intersection in positive normal direction and the distance value is stored as negative value.

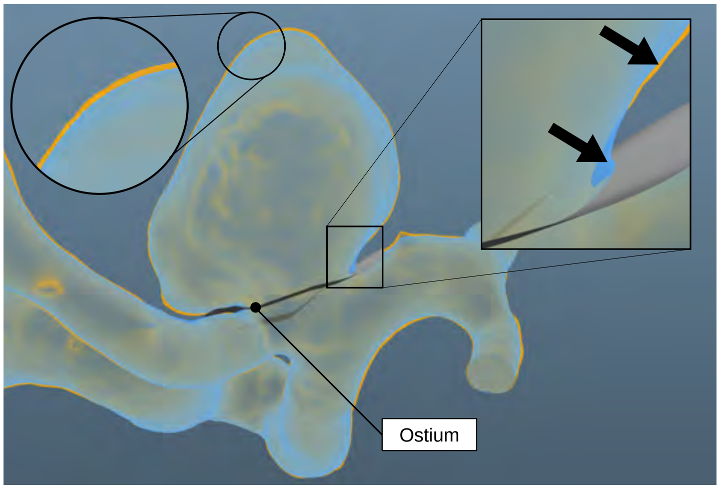


Figure 5: Depiction of Vis_B. The mesh extents become best visible at the boundary of the aneurysm (see circular inlay), which requires an interactive exploration of the 3D scene. The visualization shows a larger volume of A_{ref} at the aneurysm itself, but not at the aneurysm neck (see rectangular inlay and arrows).

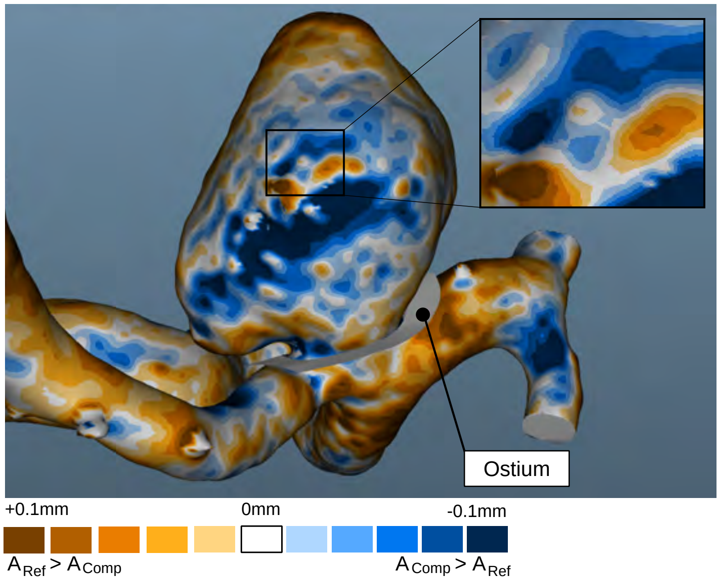


Figure 6: Depiction of visualization Vis_C. The inlay highlights the aneurysm surface.

For visual representation, we transfer the extracted distance values to the interval $[0, 1]$ since we want to store them as texture coordinates. Therefore, we clamp the original distance values to the interval $[-0.1, 0.1]$ mm and rescale them to $[0, 1]$. Thus, texture values of 0.5 are assigned to parts where the surface meshes of A_{Ref} and A_{Comp} have a distance of almost 0 mm. Finally, we employ the color map depicted in Figure 6 as texture and obtain Vis_C by using the Open Inventor Vertex Attributes module provided in MeVisLab. The color map is designed such that areas where A_{Ref} is larger than A_{Comp} are mapped to orange, whereas the quantitative distance information is provided by the hue's saturation. Blue areas indicate a larger local extent of A_{Comp}.

5 Comparative Study

In this section, we present our pipeline for a qualitative and quantitative evaluation. Afterwards, we describe our experimental setup and the user study in more detail.

5.1 A Pipeline for the Evaluation of Medical Visualizations

Based on the studies presented and discussed in Section 2, as well as discussions with statistical researchers, we created a generalizable pipeline, see Figure 7. The pipeline is reduced to the scenario of a single-factor study with one independent variable. For our application, the independent variable is the visualization with levels Vis_A, Vis_B and Vis_C. For generalization, the independent variable is provided by the medical visualization.

First, the researcher decides whether to carry out a qualitative analysis, e.g., the participants attitude towards a technique, or a quantitative analysis, e.g., to provide statistically significant results, or both. Second, user performance tasks have to be defined. Most often accuracy, e.g., the number of correct answers, and task completion time are chosen. Also the study type, i.e. between-subject design (aiming at differences between the participants) or within-subject design (aiming at the variability of a particular value for individuals in a sample), has to be chosen, which depends on the available participants. Advantageously, between-subject studies avoid learning effects and the evaluation time is reduced for each participant compared to within-subject studies. However, groups of similar participants (w.r.t. age, experience, knowledge, etc.) have to be recruited. Especially in the medical domain, these prerequisites are not easily met. On the other side, within-subject studies avoid interpersonal differences. However, they may suffer from learning or sequence effects and intrapersonal differences.

After acquiring the user performances, a test is carried out to check for a normal distribution. We can use the Shapiro-Wilk test for this purpose. This stage is a prerequisite to choose the appropriate test in the next step. We analyze whether there is a significant difference between the levels of our independent variable via an *analysis of variance (ANOVA)*. If we can assume normal distribution, i.e., the data is parametric, a one-way ANOVA (due to our single-factor study) is carried out. Otherwise, we employ the Friedmann test for the non-parametric data. If there is a significant difference between the metrics, we can examine this difference in more detail with a pairwise comparison. For example, a pair-wise comparison of the non-parametric test result for a within-subject study can be carried out with the Wilcoxon signed rank test. If no significant differences exist, we also obtained an important information. We can furthermore provide descriptive results such as the mean μ, the median m or the standard deviation σ to compare the results. Hence, interpretation of σ should take the data's distribution into account. In addition, a box plot visualization provides an important overview including information about the distribution.

Based on the infinite configuration of user studies, this pipeline is presented with no claim to completeness. However, it can be easily generalized to various application scenarios, i.e., after determining the user performance, a check for normal distribution is carried out. Next, a check for significant differences (based on one or more independent variables) and a subsequently pairwise comparison (based on one or more independent variables and within- and between-subject study design) is applied.

5.2 Experimental Setup

The whole study was realized with MeVisLab. Thus, each participant was presented with a graphical user interface (GUI), which guided the participants through the study. The user interface was created with a TabView object using hidden tabs. Each time the participant answered a question via clicking a button, the next tab was shown. At first, the TabView comprises slides for medical background information. Since all visualization techniques were implemented in MeVisLab, they could be easily integrated in the

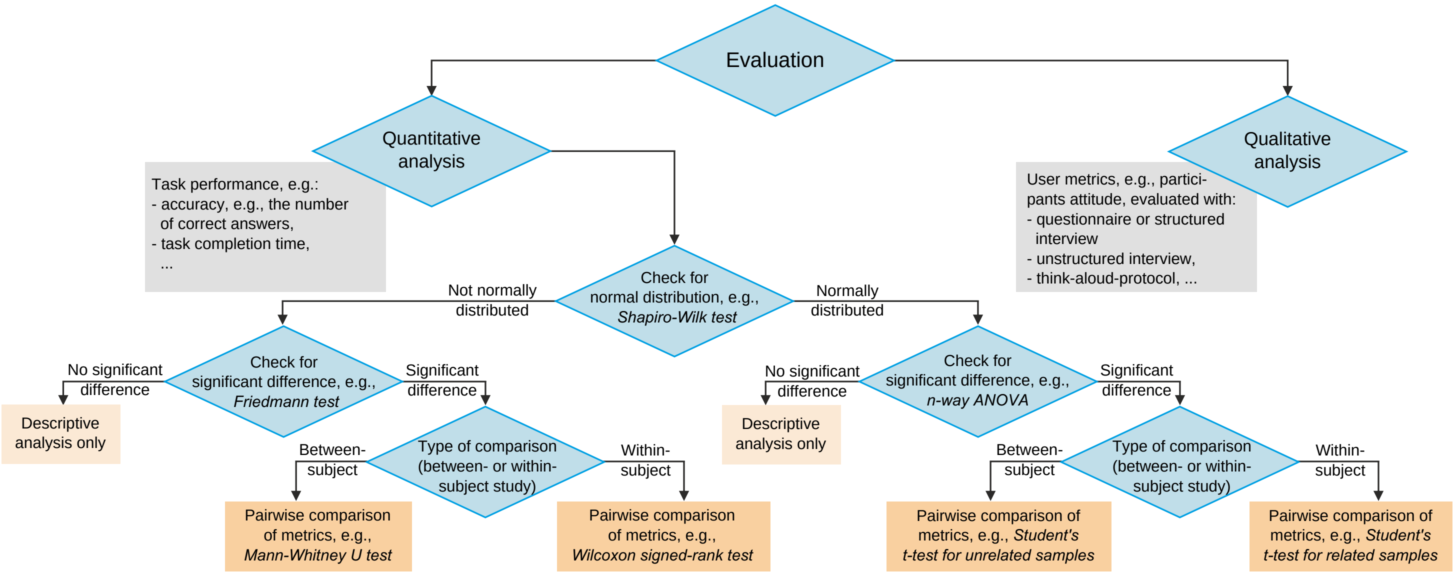

Evaluation
Quantitative analysis
Qualitative analysis
Task performance, e.g.:
- accuracy, e.g., the number of correct answers,
- task completion time,
...
User metrics, e.g., participants attitude, evaluated with:
- questionnaire or structured interview
- unstructured interview,
- think-aloud-protocol, ...
Check for normal distribution, e.g., Shapiro-Wilk test
Not normally distributed
Normally distributed
Check for significant difference, e.g., Friedmann test
No significant difference
Significant difference
Descriptive analysis only
Type of comparison (between- or within-subject study)
Between-subject
Within-subject
Pairwise comparison of metrics, e.g., Mann-Whitney U test
Pairwise comparison of metrics, e.g., Wilcoxon signed-rank test
Check for significant difference, e.g., n-way ANOVA
No significant difference
Significant difference
Descriptive analysis only
Type of comparison (between- or within-subject study)
Between-subject
Within-subject
Pairwise comparison of metrics, e.g., Student's t-test for unrelated samples
Pairwise comparison of metrics, e.g., Student's t-test for related samples

Figure 7: The proposed pipeline represented as decision tree for the qualitative and quantitative evaluation.

TabView GUI as well. Selection of visualization techniques and data sets for the participants was automatically carried out via Python scripts. Also, the logging of user inputs and time required for each task, i.e., the task completion time, were stored as text files.

5.3 Study Design

For the comparison of the 3D visualizations, we design a single-factor and within-subject study. The independent variable (i.e., the single factor) is the visualization technique which has three levels: Vis_A, Vis_B and Vis_C. The two dependent variables for each visualization are the task completion time and accuracy. Accuracy is defined as the number of correct answers, i.e., the number of right decisions whether aneurysm A_{Ref} or A_{Comp} is larger. Each experiment is carried out via a within-subject design such that each participant is confronted with each visualization technique six times. Thus, the amount of different visualization techniques shown is balanced.

Basically, we repeat the same question whether A_{ref} is larger than A_{comp} 18 times. To reduce the influence of confounder variables, e.g., training or sequence effects, we change the order of the shown visualization techniques as well as the employed patient and segmentation data with *a-priori pseudo randomization*. The pseudo randomization is listed in detail in Tab. 1, Tab. 2 and Tab. 3. For example, for the first test T_1 and the first question q_1, the user is provided with Vis_A of the data sets from patient P_1, whereas S_1 is employed for the reference aneurysm A_{Ref} and S_2 for the comparison aneurysm A_{Comp}. In general, for the i-th test T_i with questions q_1-q_{18}, each visualization Vis_A, Vis_B, and Vis_C was shown six times in the pseudo-randomized order. The patient data $P_1 - P_5$ was alternated (see Tab. 2) as well as the segmentations (see Tab. 3). Since the order of the shown data sets was important, each test is repeated for switched segmentation combinations, i.e., T_1 is identical to T_2 w.r.t. visualization technique and patient but not segmentation.

The pseudo-randomization ensures that each user evaluates different data sets with varying segmentations, i.e., the user does not see the same visualization technique with the same data sets for A_{Ref} and A_{Comp} twice. This also holds for the demonstration of visualizations during the introduction (see Sec. 5.4), where the combinations of patient data and visualization techniques were not identical to the ones used in the test.

Table 1: Pseudo-randomization for the visualizations. For the test T_i with questions q_1-q_{18}, Vis_A, Vis_B, and Vis_C were shown six times in the depicted order. Each test is repeated for switched segmentation combinations. After T_{12}, the sequence is repeated.

	q_1-q_6	q_7-q_{12}	q_{13}-q_{18}
T_1	Vis_A	Vis_B	Vis_C
T_2	Vis_A	Vis_B	Vis_C
T_3	Vis_A	Vis_C	Vis_B
...	...	...	...
T_5	Vis_B	Vis_A	Vis_C
...	...	...	...
T_7	Vis_B	Vis_C	Vis_A
...	...	...	...
T_9	Vis_C	Vis_A	Vis_B
...	...	...	...
T_{11}	Vis_C	Vis_B	Vis_A
...	...	...	...
T_{13}	Vis_A	Vis_B	Vis_C
...	...	...	...

Next to the users' choices regarding the aneurysm volumes, we logged the task completion time as well as the answers to the following questionnaire:

- the age,
- the sex,
- whether the user is familiar with 3D visualizations,
- whether the user is familiar with 3D medical image data,
- a rating for Vis_A, Vis_B and Vis_C whether the technique was suitable to analyze which aneurysm was larger, and
- a rating for Vis_A, Vis_B and Vis_C how much the user liked it.

The ratings were assessed with a 5-point Likert scale ranging from $--$ (i.e. not suitable at all or not preferable at all) to $++$ (i.e. very suitable or very preferable).

5.4 Procedure

The GUI was presented to each participant, starting with a slide for the medical background information. Afterwards, the three dif-

Table 2: Pseudo-randomization for the patient data. For the test T_i with questions q_1-q_{18}, the patient data $P_1 - P_5$ was alternated, starting with P_1 for q_1 - q_3 and the segmentations S_1-S_2, S_2-S_3, S_3-S_1 (see Tab. 3). The ordering of patients is repeated after T_{10}.

	q_1-q_3	q_4-q_6	q_7-q_9	q_{10}-q_{12}	q_{13}-q_{15}	q_{16}-q_{18}
t_1	P_1	P_2	P_3	P_4	P_5	P_1
t_2	P_1	P_2	P_3	P_4	P_5	P_1
t_3	P_2	P_3	P_4	P_5	P_1	P_2
...	...	...	...			
t_5	P_3	P_4	P_5	P_1	P_2	P_3
...	...	...	...			
t_7	P_4	P_5	P_1	P_2	P_3	P_4
...	...	...	...			
t_9	P_5	P_1	P_2	P_3	P_4	P_5
...	...	...	...			
t_{11}	P_1	P_2	P_3	P_4	P_5	P_1
...	...	...	...	...	...	...

ferent visualizations Vis_A, Vis_B, and Vis_C were shown. Each of the visualizations as well as the interaction, e.g., zooming and rotating, was explained in detail by the supervisor. The user was also encouraged to explore the scene and get familiar with the user interface for 3D exploration provided by MeVisLab. The test number t_i was assigned to the i-th user. The user had to answer 18 questions q_1-q_{18} and decide, which aneurysm possess the larger volume. Finally, the users answered the questionnaire.

6 Results

This section describes the participants and lists the results of the user study including quantitative and descriptive analyses, based on our evaluation pipeline (recall Fig. 7). Afterwards, the qualitative subjective ratings w.r.t. suitability and preferability are discussed.

6.1 Participants

The participants were recruited from visitors of the *Long Night of Sciences*. During this event, scientific institutes show experiments and tests to the general public. The majority of our participants were from the university's computer science and medical engineering departments. As a result, we were able to conduct a user study with 34 participants comprising 5 female and 29 male users, aging from 16 - 66 years. When asked if the users have experiences with medical visualizations, 10 users declined and 24 affirmed. Regarding the experience with 3D visualizations, eight users stated they have no experience. We did not include domain experts or prospective users since we were only interested in a perceptual evaluation of volume change. Hence, no medical knowledge was required.

Table 3: Pseudo-randomization for the segmentations. For the test T_i with questions q_1-q_{18} two different segmentations of the same patient were employed. For example, S_1-S_2 indicates segmentation S_1 for A_{Ref} and segmentation S_2 for A_{Comp}. Since the order of the shown data sets was important, the order of segmentations is reversed for odd tests.

	$q_{1,4,7,10,13,16}$	$q_{2,5,8,11,14,17}$	$q_{3,6,9,12,15,18}$
T_1	S_1-S_2	S_2-S_3	S_3-S_1
T_2	S_2-S_1	S_3-S_2	S_1-S_3
T_3	S_1-S_2	S_2-S_3	S_3-S_1
...	...	...	...

Table 4: Data from the user study. For each user U_1-U_{34}, the number of correct answers for Vis_A, Vis_B and Vis_C is extracted. This value ranges from 0 to 6, since each participant was confronted with each technique six times. Also, for each user the average time t_A, t_B and t_C (provided in seconds) to answer a question is collected.

	Correct answers			Average required time		
	Vis_A	Vis_B	Vis_C	t_A	t_B	t_C
U_1	5	2	5	22 s	20.17 s	17.67 s
...	...	...	...	...	...	...
U_{34}	5	4	5	17.83 s	10.67 s	15.33 s

6.2 Evaluation

The data collection provided by the conducted user study is listed in Table 4. The participants' answers form the set of observations for Vis_A, Vis_B and Vis_C. We also collect the set of averaged task completion times t_A, t_B, and t_C, each participant needed for Vis_A, Vis_B and Vis_C. All statistical tests were carried out with SPSS 22.0 (IBM, New York, USA). Our statistical analysis comprises three stages (recall Fig. 7):

1. We determine whether there is a significant difference between the visualizations w.r.t. accuracy.

2. In case the visualizations are significantly different, we further analyze which visualization technique is best suited w.r.t. accuracy and task completion time by pairwise comparison.

3. Finally, we provide a descriptive analysis.

6.2.1 Statistical Analysis Regarding the Accuracy

First Stage. The first analysis stage determines whether there is significant difference between the three visualization techniques w.r.t. the amount of right answers, recall Tab. 4. Box plots for the accuracy for Vis_A, Vis_B and Vis_C are provided in Figure 8. Initially, we employ the *Shapiro-Wilk* test separately for Vis_A, Vis_B and Vis_C to determine whether the amount of right answers is normally distributed. Hence, the null-hypothesis H_0 of the test states a normal distribution of the random variable:

$$H_0 : \textit{The random variable is normally distributed.}$$

The Shapiro-Wilk test yields the following significance levels:

- 0.003 for Vis_A, and
- 0.037 for Vis_B, and
- 0.000 for Vis_C.

Since H_0 is rejected, if the significance level is smaller than 0.05, the accuracy significantly deviates from a normal distribution for each visualization technique. The next step comprises the analysis, whether the visualization techniques are significantly different. We chose the *Friedmann* test, since this test provides an ANOVA for random variables that are not normally distributed. We define the hypothesis:

$$H_0 : \textit{All visualization techniques achieve similar results.}$$

Advantageously, the Friedmann test is based on ranks and not the actual scores. The Friedman test reveals that the accuracies significantly differ for the three visualizations, with $\chi^2(2) = 25.382$, $p < .05$, and the hypothesis H_0 must be rejected.

Second Stage. In the second analysis stage, we compare the visualization techniques to identify the best one w.r.t. accuracy. Based on the previous results, i.e., the amounts of right answers are not normally distributed and are significantly different, we carry out a pair-wise comparison of the visualizations. Since we deal with non-parametric data, we apply the *Wilcoxon signed-rank* test for Vis_A, Vis_B and Vis_C. A correction with the Bonferroni procedure [Sha95] was applied, since we carry out multiple tests on the participants' responses. Thus, all effects are reported at a .0167 level of significance, i.e., a third of 0.05. The amount of correct answers were significantly higher for Vis_A ($m = 4.5$) than for Vis_B ($m = 3.0$), $U = -3.76, p < .0167$, where m denotes the median. Also, the amount of correct answers was significantly higher for Vis_C ($m = 5.0$) than for Vis_B ($m = 3.0$), $U = -4.07, p < .0167$. However, there was no significant difference between Vis_A ($m = 4.5$) and Vis_C ($m = 5.0$), $U = 0.95, p = .354$. The resulting box plots for Vis_A, Vis_B and Vis_C are provided in Figure 8.

Since Vis_B significantly differs from Vis_A and Vis_C, we analyzed how it competes with guessing, where guessing would result in three correct answers. Hence, a Wilcoxon signed rank test yields a significant difference ($U = -2,094, p < .05$ with $\mu_{Vis_B} {<} \mu_{guessing}$). Thus, Vis_B may systematically influence the users to provide wrong answers.

Third Stage. When using Vis_C ($\mu = 4.47, \sigma = 1.16$) and Vis_A ($\mu = 4.06, \sigma = 1.67$), the participants achieved a higher accuracy than with Vis_B ($\mu = 2.41, \sigma = 1.52$). Comparison of the mean values of Vis_A and Vis_C indicates the superiority of Vis_C.

6.2.2 Statistical Analysis Regarding the Required Time

For each visualization technique, the task completion time was logged. We averaged the task completion time for each question, i.e., we extract the average time t_A, t_B and t_C required by the users for a single question using Vis_A, Vis_B, or Vis_C, respectively (recall Tab. 4). The boxplots are depicted in Figure 9. Similar to the previous analysis, we first determine whether there is a statistically significant difference between t_A, t_B and t_C. We employ the *Shapiro-Wilk* test to determine whether the required times are normally distributed yielding the following significance levels:

- 0.029 for t_A, and
- 0.007 for t_B, and
- 0.006 for t_C.

Hence, all three variables significantly deviate from a normal distribution (p < 0.05).

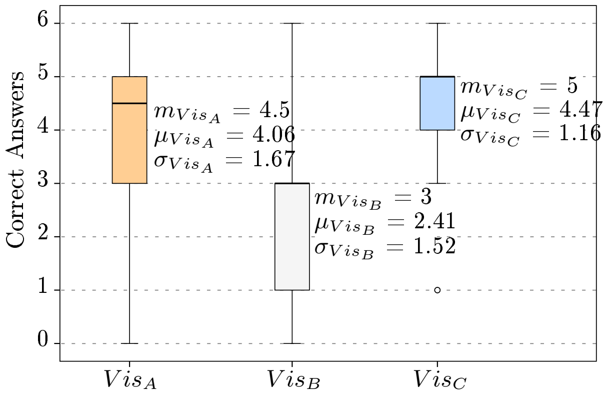


Figure 8: Box plots of the accuracy for Vis_A, Vis_B and Vis_C including the median m, the mean μ and the standard deviation σ.

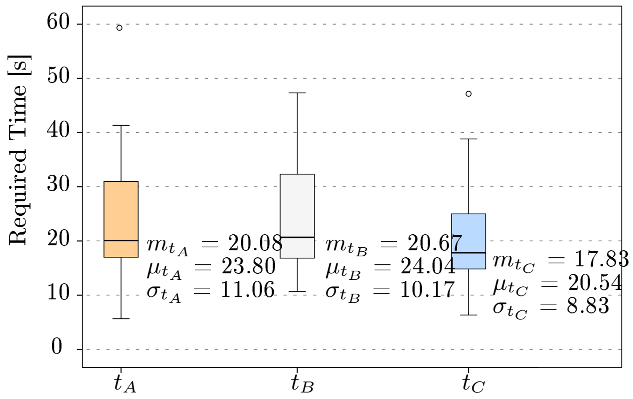


Figure 9: Box plots of the averaged task completion times t_A, t_B, and t_C including the median m, the mean μ and the standard deviation σ.

As proposed for statistical analysis of Vis_A, Vis_B and Vis_C w.r.t. the accuracy, the second stage determines whether t_A, t_B and t_C are significantly different. Therefore, we carry out the Friedmann test, since this test provides an ANOVA for random variables that are not normally distributed. The corresponding null-hypothesis is:

H_0 : *The task completion time differs for Vis_A, Vis_B and Vis_C.*

As a result, the Friedman test reveals no significant difference, i.e., $\chi^2(2) = 2.8$, and $p > 0.05$. Thus, H_0 cannot be rejected.

Second Stage. Since no statistically significant difference could be shown by the Friedmann test, we do not carry out a pairwise comparison of the task completion times.

Third Stage. Comparing the box plots and test statistics of t_A, t_B and t_C, the participants performed the tasks in average faster with Vis_C ($\mu = 20.54, \sigma = 8.83$) compared to Vis_A ($\mu = 23.80, \sigma = 11.06$) and Vis_B ($\mu = 24.04, \sigma = 10.17$). Comparing the mean values of t_A and t_B, the users required more time to fulfill the tasks with Vis_B.

6.2.3 Qualitative Evaluation of Suitability and Preferability

When analyzing the suitability and preferability ratings, the same trends are reflected, see Figure 10. Furthermore, the mode value, i.e., the answer ($--, -, 0, +, ++$) that was given most often for each question, as well as the amount of users that provide answer $++$ and $+$ is provided. Hence, users mostly rated Vis_C with $++$ for suitability as well as preferability, Vis_A with $+$ for suitability as well as preferability and Vis_B with $-$ for suitability as well as preferability. The amount of users rating Vis_C as suitable and very suitable (i.e., answers are $+$ or $++$) was highest with 27, followed by 21 for Vis_A and 9 for Vis_B. Similarly, the amount of users rating Vis_C as preferable and very preferable (i.e., answers are $+$ or $++$) was highest with 29, followed by 16 for Vis_A and 11 for Vis_B.

7 Discussion

The statistical analysis revealed a significant difference of Vis_A, Vis_B and Vis_C w.r.t. accuracy. The pair-wise comparison identifies Vis_B as poorest choice. It does not only achieve lower mean values compared to Vis_A and Vis_C, but significantly differs from both as well. Vis_C is not statistically significant different from Vis_A, however, due to the higher mean values compared to Vis_A,

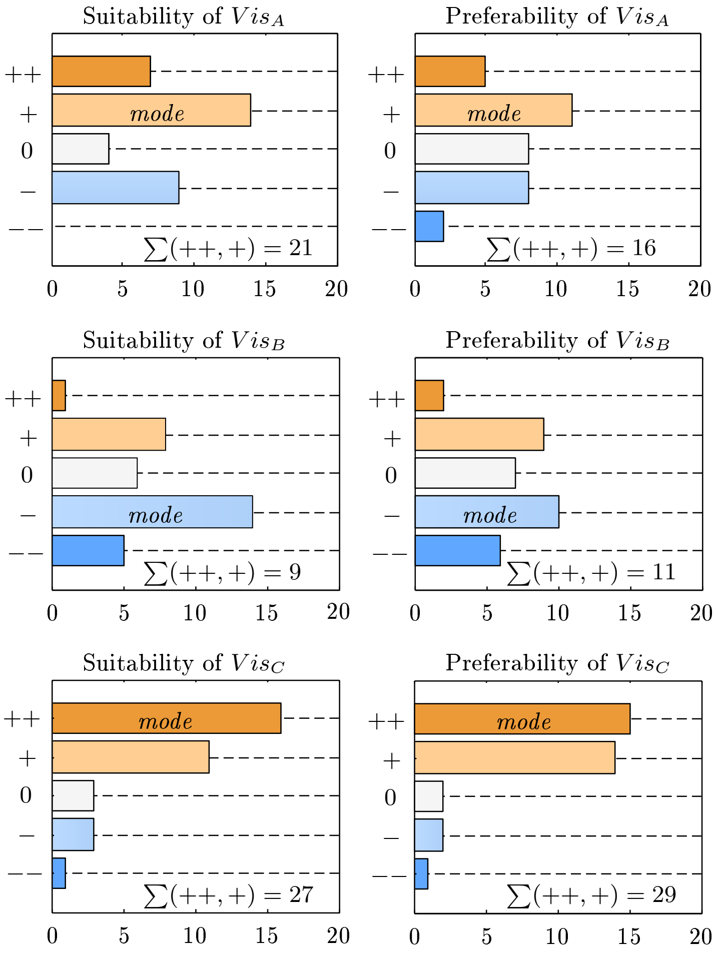


Figure 10: Evaluation results of the participants regarding suitability and preferability of Vis_A, Vis_B, and Vis_C. The mode value, i.e., the answer that was given most often for each question, is marked. Furthermore, the sum of answers ++ and + is provided.

it is declared as the best visualization to compare the volume of two aneurysms. A possible conclusion might be that a derived quantity, i.e., the distance, improves the identification of the larger aneurysm. Furthermore, mean and median values of t_C were smaller than the values of t_A and t_B. Although no significant difference occurred, these test results rate Vis_C as best visualization w.r.t. task completion time.

Remarkably, Vis_B even achieved a lower success rate than guessing. This is interpreted as indication that the users did not fully understand the design of Vis_B and that Vis_B is very inappropriate for comparison of surfaces. We assume that users wrongly interpret the ghosting view and thus, do not focus on the border areas but instead on areas with surface normals parallel to the current viewing direction. These areas are pre-dominantly color-coded in cyan, since the A_{Comp} aneurysm is always drawn after the orange A_{Ref} aneurysm.

When analyzing the suitability and preferability ratings, we found overwhelming preference for Vis_A and Vis_C over Vis_B which further indicates the inappropriateness of the latter. There was also a small trend towards preferring Vis_C over Vis_A, identifying Vis_C as favorite visualization.

8 Conclusion

Researchers involved in medical applications are often confronted with visualization techniques, which are rather difficult to evaluate. Many times, medical visualization papers lack a quantitative evaluation at all. With our proposed user study, a pipeline was presented, which allows the comparative evaluation for three different visualization techniques for the specific application of cerebral aneurysm volume assessment. With focus on accuracy and task completion time, this concept can be easily applied to various scenarios to support qualitative findings with quantitative results.

For the evaluation of the aneurysm volume, the visualization should be reduced to basic information, i.e., no ghosted view techniques should be employed. Providing a color-coded surface visualization with quantitative distance information such as our new visualization technique Vis_C, helps the users to decide which aneurysm exhibits the largest volume. This was reflected by a statistically significant higher accuracy, a smaller task completion time as well as a better user rating.

For future work, different approaches can be pursued. The visualizations can be improved, for example by including depth cues such as ambient occlusion. From the statistical point of view, a systematic analysis of the influence of the volume change could be carried out. Hence, a visualization technique may be well-suited for the depiction of large volume changes, but rather improperly suited for small volume changes with a second visualization technique exhibiting the opposite behavior. Finally, we chose the employed colors to prevent false interpretations due to red-green color blindness. In future, different color blindness types should be considered and assessed with the questionnaire.

Acknowledgements We warmly thank Alexandra Baer and Maria Luz for fruitful discussions and guidance for the evaluation. We also thank the Institute for Neuroradiology, University of Magdeburg, for providing the medical image data. This work was partially funded by the Federal Ministry of Education and Research within the Forschungscampus *STIMULATE* (grant number '13GW0095A').

References

[BBF$^+$11] Stef Busking, Charl P Botha, Luca Ferrarini, Julien Milles, and Frits H Post. Image-based rendering of intersecting surfaces for dynamic comparative visualization. *The Visual Computer*, 27(5):347–363, 2011.

[BCFW08] Dirk Bartz, Douglas Cunningham, Jan Fischer, and Christian Wallraven. The role of perception for computer graphics. *Eurographics state-of-the-art-reports*, pages 65–86, 2008.

[BFLC04] Katja Bühler, Petr Felkel, and Alexandra La Cruz. *Geometric methods for vessel visualization and quantification–a survey*. Springer, 2004.

[BGCP11] A. Baer, R. Gasteiger, D. Cunningham, and B. Preim. Perceptual evaluation of ghosted view techniques for the exploration of vascular structures and embedded flow. *Computer Graphics Forum*, 30(3):811–820, 2011.

[BGP$^+$11] M. Borkin, K. Gajos, A. Peters, D. Mitsouras, S. Melchionna, F. Rybicki, C. Feldman, and H. Pfister. Evaluation of artery visualizations for heart disease diagnosis. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2479–2488, 2011.

[BRB$^+$15] P. Berg, C. Roloff, O. Beuing, S. Voss, S. Sugiyama, N. Aristokleous, and et al. The Computational Fluid Dynamics Rupture Challenge 2013 – Phase II: variability of hemodynamic simulations in two intracranial aneurysms. *Journal of Biomechanical Engineering*, 137(12):121008/1–13, 2015.

[CCA+05] J. R. Cebral, M. A. Castro, S. Appanaboyina, C. M. Putman, D. Millan, and A. F. Frangi. Efficient pipeline for image-based patient-specific analysis of cerebral aneurysm hemodynamics: technique and sensitivity. *IEEE Transactions on Medical Imaging*, 24(4):457–467, 2005.

[CFM+13] Robert Carnecky, Raphael Fuchs, Stephanie Mehl, Yun Jang, and Ronald Peikert. Smart transparency for illustrative visualization of complex flow surfaces. *IEEE Transactions on Visualization and Computer Graphics*, 19(5):838–851, 2013.

[CSP10] J. R. Cebral, M. Sheridan, and C. M. Putman. Hemodynamics and bleb formation in intracranial aneurysms. *American Journal of Neuroradiology,*, 31(2):304–310, 2010.

[DRN+15] Jose Díaz, Timo Ropinski, Isabel Navazo, Enrico Gobbetti, and Pere-Pau Vázquez. An experimental study on the effects of shading in 3d perception of volumetric models. *The Visual Computer*, pages 1–15, 2015.

[GBNP15] S. Glaßer, P. Berg, M. Neugebauer, and B. Preim. Reconstruction of 3d surface meshes for blood flow simulations of intracranial aneurysms. In *Proc. of Computer and Robotic Assisted Surgery (CURAC)*, pages 163–168, 2015.

[GLH+14] Sylvia Glaßer, Kai Lawonn, Thomas Hoffmann, Martin Skalej, and Bernhard Preim. Combined visualization of wall thickness and wall shear stress for the evaluation of aneurysms. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, pages 2506–2515, 2014.

[GLR+09] Arjan J Geers, Ignacio Larrabide, AG Radaelli, Hrvoje Bogunovic, HAFG Van Andel, CB Majoie, and Alejandro F Frangi. Reproducibility of image-based computational hemodynamics in intracranial aneurysms: comparison of CTA and 3DRA. In *Proc. of IEEE Symposium on Biomedical Imaging: From Nano to Macro (ISBI)*, pages 610–613, 2009.

[GNKP10] R. Gasteiger, M. Neugebauer, C. Kubisch, and B. Preim. Visualization of cerebral aneurysms with embedded blood flow information. In *Proc. of the Eurographics Workshop on Visual Computing for Biology and Medicine (VCBM)*, pages 25–32, 2010.

[GR04] Gevorg Grigoryan and Penny Rheingans. Point-based probabilistic surfaces to show surface uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 10(5):564–573, 2004.

[GSK+15] Alexander Geurts, Georgios Sakas, Arjan Kuijper, Meike Becker, and Tatiana von Landesberger. Visual comparison of 3d medical image segmentation algorithms based on statistical shape models. In *Digital Human Modeling. Applications in Health, Safety, Ergonomics and Risk Management: Ergonomics and Health*, pages 336–344. Springer, 2015.

[IIC+13] Tobias Isenberg, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Moller. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013.

[KHI+03] Robert Kosara, Christopher G Healey, Victoria Interrante, David H Laidlaw, and Colin Ware. User studies: why, how, and when? *IEEE Computer Graphics and Applications*, 23(4):20–25, 2003.

[LABFL09] David Lesage, Elsa D Angelini, Isabelle Bloch, and Gareth Funka-Lea. A review of 3D vessel lumen segmentation techniques: models, features and extraction schemes. *Medical Image Analysis*, 13(6):819–845, 2009.

[LEBB09] R. R. Lall, Christopher S. Eddleman, Bernard R. Bendok, and H. Hunt Batjer. Unruptured intracranial aneurysms and the assessment of rupture risk based on anatomical and morphological factors: sifting through the sands of data. *Neurosurgical Focus*, 26(5):E2, 2009.

[MMNG15] Haichao Miao, Gabriel Mistelbauer, Christian Našel, and M Eduard Gröller. Cowradar: Visual quantification of the circle of willis in stroke patients. In *Proc. of the Eurographics Workshop on Visual Computing for Biology and Medicine (VCBM)*, pages 1–10, 2015.

[NLBP13] Mathias Neugebauer, Kai Lawonn, Oliver Beuing, and Bernhard Preim. Automatic generation of anatomic characteristics from cerebral aneurysm surface models. *International Journal of Computer Assisted Radiology and Surgery*, 8(2):279–289, 2013.

[PBC+16] Bernhard Preim, Alexandra Baer, Douglas Cunningham, Tobias Isenberg, and Timo Ropinski. A survey of perceptually motivated 3d visualization of medical image data. *Computer Graphics Forum*, 35(3):501–525, 2016.

[PH11] Kai Pöthkow and Hans-Christian Hege. Positional uncertainty of isocontours: Condition analysis and probabilistic measures. *IEEE Transactions on Visualization and Computer Graphics*, 17(10):1393–1406, 2011.

[PO08] Bernhard Preim and Steffen Oeltze. 3d visualization of vasculature: an overview. In *Visualization in medicine and life sciences*, pages 39–59. Springer, 2008.

[PRJ12] Kristin Potter, Paul Rosen, and Chris R Johnson. From quantification to visualization: a taxonomy of uncertainty visualization approaches. In A Dienstfrey and R Boisvert, editors, *Uncertainty Quantification in Scientific Computing*, volume 377, pages 226–249. Springer, 2012.

[Sha95] Juliet Popper Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–584, 1995.

[SOBP07] Christian Schumann, Steffen Oeltze, Ragnar Bade, and Bernhard Preim. Model-free surface visualization of vascular trees. In *Proc. of IEEE/Eurographics Symposium on Visualization (EuroVis)*, pages 283–290, 2007.

[Wie03] D. O. Wiebers. Unruptured intracranial aneurysms: natural history, clinical outcome, and risks of surgical and endovascular treatment. *The Lancet*, (9378):103–110, 2003.

[WT05] C. Weigle and R. M. Taylor. Visualizing intersecting surfaces with nested-surface techniques. In *Proc. of IEEE Visualization*, pages 503–510, 2005.

[WvdSAR07] M. J. Wermer, I. C. van der Schaaf, A. Algra, and G. J. Rinke. Risk of rupture of unruptured intracranial aneurysms in relation to patient and aneurysm characteristics: an updated meta-analysis. *Stroke*, 38(4):1404–1410, 2007.

[XNT+11] J. Xiang, S. K. Natarajan, M. Tremmel, D. Ma, J. Mocco, L. N. Hopkins, A. H. Siddiqui, E. I. Levy, and H. Meng. Hemodynamic-morphologic discriminants for intracranial aneurysm rupture. *Stroke*, 42(1):144–152, 2011.