Guidelines for Quantitative Evaluation of Medical Visualizations on the Example of 3D Aneurysm Surface Comparisons

P. Saalfeld, M. Luz, P. Berg, B. Preim, S. Saalfeld

DOI: 10.1111/cgf.13262

Guidelines for Quantitative Evaluation of Medical Visualizations on the Example of 3D Aneurysm Surface Comparisons

P. Saalfeld[1], M. Luz[1], P. Berg[2], B. Preim[1] and S. Saalfeld[1]

[1]Department of Simulation and Graphics, Otto-von-Guericke University, Magdeburg, Germany
patrick@isg.cs.uni-magdeburg.de, maria.luz@ovgu.de, bernhard@isg.cs.uni-magdeburg.de, sylvia.saalfeld@ovgu.de
[2]Department of Fluid Dynamics and Technical Flows, Otto-von-Guericke University, Magdeburg, Germany
philipp.berg@ovgu.de

Abstract

Medical visualizations are highly adapted to a specific medical application scenario. Therefore, many researchers conduct qualitative evaluations with a low number of physicians or medical experts to assess the benefits of their visualization technique. Although this type of research has advantages, it is difficult to reproduce and can be subjectively biased. This makes it problematic to quantify the benefits of a new visualization technique. Quantitative evaluation can objectify research and help bringing new visualization techniques into clinical practice. To support researchers, we present guidelines to quantitatively evaluate medical visualizations, considering specific characteristics and difficulties. We demonstrate the adaptation of these guidelines on the example of comparative aneurysm surface visualizations. We developed three visualization techniques to compare aneurysm volumes. The visualization techniques depict two similar, but not identical aneurysm surface meshes. In a user study with 34 participants and five aneurysm data sets, we assessed objective measures (accuracy and required time) and subjective ratings (suitability and likeability). The provided guidelines and presentation of different stages of the evaluation allow for an easy adaptation to other application areas of medical visualization.

Keywords: evaluation, medical visualization, aneurysm surface comparison

ACM CCS: I.3.3 [Computer Graphics]: Picture/ImageGeneration and Display Algorithms, G.3 Probability and Statistics Experimental Design J.2 Physical Sciences and Engineering Mathematics and Statistics

1. Introduction

Medical visualizations are developed to support the in-depth understanding of diagnostic processes, therapeutic decisions and to satisfy intra-operative information needs. Evaluation is mandatory to assess existing visualization techniques, develop new ones, answer research questions and generate and verify postulated hypotheses. Here, a wide variety of evaluation strategies exists. Since the visualization techniques are highly adapted to the specific medical application scenario, prior knowledge is often required, which narrows the range of eligible participants. As a result, many researchers conduct qualitative evaluations with a low number of medical experts to assess the benefits of their visualization technique. However, the acquired results are difficult to reproduce. Furthermore, the medical experts usually are cooperation partners and co-authors of the presented work, where a subjective bias is hardly avoidable. Hence, quantitative evaluation can objectify research, provide additional information and determine whether a statistically significant difference is achieved.

In this paper, we present guidelines for the statistical evaluation of medical visualizations based on the example of comparative aneurysm surface views. We discuss possible study designs and list common measurable properties to assess users' objective and subjective performance. The subsequent analysis allows for determination of statistical significance.

Our medical application scenario covers intracranial aneurysms. The segmentation of such vessel pathologies is an important research area. To create reproducible results and 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]. 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 extraction of morphological parameters [LEBB09] or the simulation of the internal blood flow [BRB*15]. 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. During the segmentation process, the medical expert requires feedback on how parameters influence the segmentation results, since small parameter adjustments may induce enormous changes on the surface mesh. To guide the clinical expert through the process, we developed three different comparative visualization techniques to show surface mesh variations.

Our quantitative evaluation determines the most suitable visualization technique to assess changes in the aneurysm volumes. Here, we consider objective measures and subjective ratings. The visualization techniques are applied to five cerebral aneurysms, each approximated with three slightly different surface meshes.

This work is an extension of our previous work [GSB*16]. We use the application scenario of cerebral aneurysms to provide three techniques for the visualization of two similar but not identical aneurysm surface meshes, which mutually penetrate and overlap. The additional contributions of this paper are:

- We present comprehensive guidelines to quantitatively evaluate medical visualizations, considering specific characteristics and difficulties. Here, we provide instructions for computer scientists and engineers to carry out statistical evaluation.
- These guidelines are represented as a decision tree, comprising the most common statistical tests. The tree can be used as guidance leading researchers from their research question to the choice of a matching statistical test for the desired quantitative evaluation.
- In addition to the identification of the best suited visualization technique regarding accuracy and required time, we also carry out a quantitative evaluation of user subjective ratings, yielding statistically significant results.
- Finally, we evaluate whether the participants' experience with medical visualizations has a significant influence on their accuracy and required time to decide which aneurysm possesses the larger volume.

2. Related Work

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]. Lam *et al.* [LBI*12] introduced an in-depth discussion of seven evaluation scenarios for information visualization, which are subdivided in scenarios for understanding data analysis processes and in scenarios for visualization evaluation. Their approach focused on evaluation goals and questions that guide the users to select appropriate methods based on the provided context within the different scenarios. Our proposed pipeline can be categorized into the evaluation of user performance, evaluation of visualization type, as well as evaluation of visual data analysis and reasoning. We chose the quantitative statistical evaluation as a goal and provide detailed information as well as the required statistical tests to achieve it.

Isenberg *et al.* [IIC*13] presented a systematic review of the evaluation practices in visualization. They employed several evaluation categories and concluded that the *Qualitative Result Inspection* was most often used by all reviewed papers. Further emphasis on the evaluation of algorithmic performance as well as an increasing trend in the evaluation for user experience and user performance were reported.

Examples for this quantitative trend in medical visualizations are user studies performed by Gasteiger *et al.* [GNKP10] and Baer *et al.* [BGCP11]. Gasteiger *et al.* evaluated an aneurysm visualization based on the participant's grade of satisfaction w.r.t. depth perception, spatial relationships, flow perception and surface shape. Subsequently, Baer *et al.* [BGCP11] compared this visualization technique against two others and were able to determine statistically significant differences for the visualizations. Borkin *et al.* [BGP*11] determined which visualization technique of the endothelial shear stress of coronary arteries is best suited. The 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. The survey by Preim *et al.* [PBC*16] presents perception-based evaluations of medical visualization techniques focusing on shape and depth cues. They proposed to design studies in such a way that a broad range of users can participate by creating tasks that are solvable with general visual perception abilities. It provides essential aspects of perceptual experiment methods as well as a discussion of the type and setting of an evaluation, stimuli, participants, tasks and major results for selected medical visualization techniques. In contrast, the presented approach focuses more on detailed information about the required tests for a quantitative statistical evaluation, but also provides general information about study design and experimental setup choices.

Visualizations of vessels are often depicted as 3D surfaces due to their complex and patient-individual shape [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 morphological analyses and subsequent Computational Fluid Dynamics (CFD) simulations, 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 colour-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.

One of our visualizations 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 colour-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 volumes. Therefore, we want to reduce the visual complexity and choose the concepts provided by Busking *et al.* [BBF*11] as inspiration for one of our visualization techniques (Vis_C).

Our visualization techniques show different segmentation results from the same patient, which can also be 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 iso-surfaces with uncertainties. Their approach employs colour-coding, glyphs and direct volume rendering.

The presented approach only covers a specific part of a medical application scenario and explains which statistical test can be adapted to evaluate the medical visualization. In the longer term, medical visualization aims at the support of medical decision making. For example, Lang *et al.* [LRHea05] reported a change of operation planning due to the influence of computer-assisted risk analysis.

3. Comparative Visualization of Cerebral Aneurysms

This section presents the aneurysm image data, the segmentation process and the three visualization techniques Vis_A, Vis_B and Vis_C.

3.1. Cerebral aneurysm image data and image processing

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. To avoid unnecessary treatment, rupture risk assessment is an active clinical research area.

In clinical practice, rupture risk factors mainly comprise the aneurysm's morphology and whether the aneurysm is asymptomatic or symptomatic [WvdSAR07]. Hence, the extraction of aneurysm surface meshes provides 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]. Again, a patient-specific surface mesh is the prerequisite for volume grid extraction and a subsequent CFD simulation.

For the 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 (P_1–P_5) were reconstructed using the Hounsfield Units (HU) setting and three different image characteristics: smooth, normal and sharp [BSV*17]. The HU kernel is recommended for quantitative measurements. The sharp setting maximizes spatial resolution but yields increased noise levels, whereas the smooth setting reduces artifacts as well as the spatial resolution. A compromise between smooth and sharp is provided by the normal setting [syn16]. 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 internal carotid artery and one at the bifurcation of the middle cerebral artery. Their size varied from 2.5 to 11.2 mm (mean size). All patients were treated with endovascular coiling.

Reconstructing the RA data, P_1–P_5 with the three different reconstruction modes yields three DSA data sets for each patient. Aneurysm segmentation was carried out via thresholding [GBNP15]. The segmentation and surface mesh generation was performed in MeVisLab 2.7 (MeVis Medical Solutions AG, Bremen, Germany). To provide a visual separation between parent vessel and aneurysm, we extracted an ostium for each patient using Blender 2.74 (Blender Foundation, Amsterdam, the Netherlands). The ostia were extruded to create ruff-like structures in order to support the participants and the evaluation of the aneurysm size. The extraction of surface meshes and ostia is described in more detail in [GSB*16]. Figure 1 illustrates the aneurysm surface meshes for P_1–P_5 as well as surface meshes for a single patient based on the three reconstruction modes.

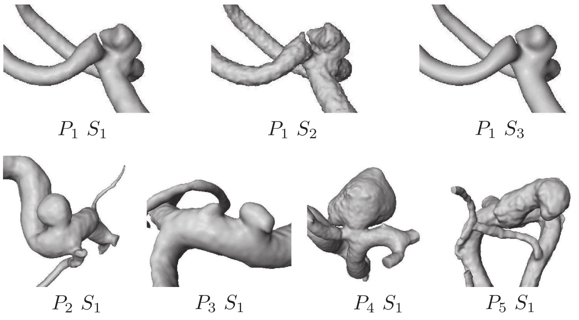

Figure 1: *Depiction of aneurysm surface meshes. For patient P_1, the three resulting segmentations S_1, S_2 and S_3 based on the three reconstruction modes (HU normal, HU sharp and HU smooth) are shown (top). Surface meshes of the remaining patients P_2–P_5 reconstructed with HU normal are visualized (bottom).*

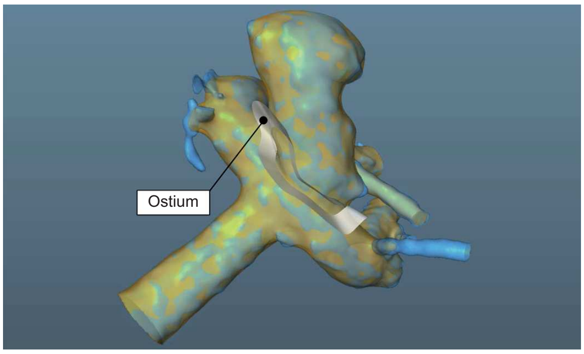

Figure 2: *Depiction of the iso-surface view Vis_A. In case the surface mesh of A_{Ref} exceeds the surface mesh of A_{Comp}, the orange surface becomes visible. Otherwise, the cyan mesh is visible. The ruff-like structure provides information about the ostium.*

3.2. 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 colour-coded map surface view Vis_C. Vis_A and Vis_B show 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.

3.2.1. *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 [$RGB = (1, 0.33, 0)$], and for A_{Comp} a cyan [$RGB = (0.33, 0.66, 1)$] transparent surface mesh is simultaneously visualized with opacity values of 0.5 (see Figure 2). The colour-coding uses complementary colours and accounts for red–green colour blindness. Beyond mesh extraction, no further preprocessing is required.

3.2.2. *The boundary-enhanced view – Vis_B*

The second visualization technique Vis_B (see Figure 3) 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 colour-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 $\vec{n}$ and the viewing vector $\vec{v}$:

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

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 same colours are used for Vis_A and Vis_B. The visualization technique is realized in MeVisLab using the Open Inventor vertex and fragment shader modules where the user can directly provide shader code as input.

3.2.3. *The 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. For visual representation, we normalize 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 (a well suited range for small structures such as cerebral aneurysms) 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 colour map depicted in Figure 4 as texture and obtain Vis_C by using the Open Inventor Vertex Attributes module provided in MeVisLab. The colour map is based on the chosen colours for Vis_A and Vis_B. It is designed such that areas where A_{Ref} is larger than A_{Comp} are mapped to dark orange, whereas the quantitative distance information is provided by the hue's saturation. Blue areas indicate a larger local extent of A_{Comp}.

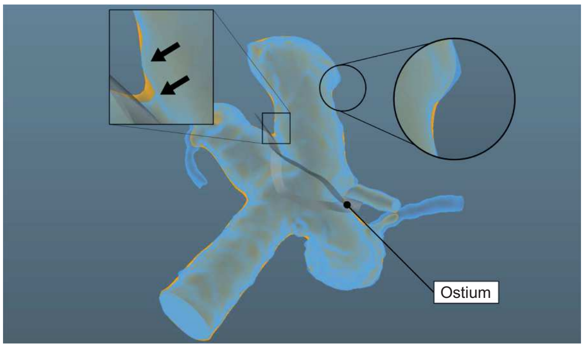

Figure 3: *Depiction of Vis_B. The mesh extents become best visible at the boundary of the aneurysm (see circular inlay), which requires an interactive exploration of the 3D scene. The visualization shows a larger aneurysm neck of A_{Ref} (see rectangular inlay and arrows).*

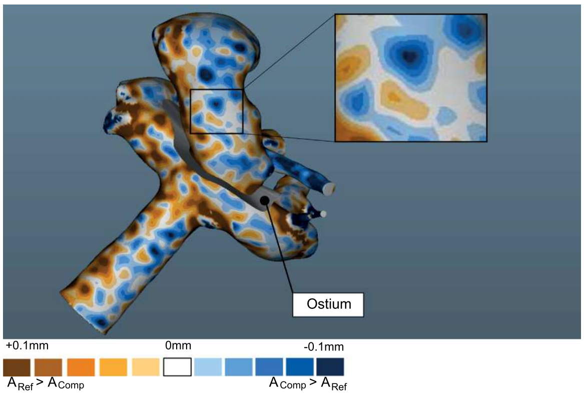

Figure 4: *Depiction of visualization Vis_C. Similar to a relief map, colour-coding provides information whether A_{Ref} or A_{Comp} is larger. Hence, the colour saturation provides quantitative information.*

4. Guidelines for the Evaluation of Medical Visualization

Based on previous evaluation projects, discussions with statistical experts and the studies presented in Section 2, we derive guidelines for the evaluation of medical visualizations. These guidelines are summarized as a decision tree with several stages, see Figure 5, with focus on inferential statistics. These stages are described in more detail in the following.

The most general subdivision of methods is the distinction between quantitative and qualitative methods (Figure 5, Stage 1). While the former allows an analysis of *measurable properties*, the latter investigates *phenomena*. Important to note is that *user subjective ratings*, e.g. assessed with a Likert scale, are measurable properties as well. Quantitative evaluation methods can be applied to measurable properties to determine whether statistically significant findings can be extracted. On the other hand, qualitative evaluation is the right choice for explorative research questions to generate hypotheses as well as to provide basic information for a new application area. For example, if a new medical visualization should be developed, qualitative evaluation can be applied to determine the requirements for the novel visualization. Also, the decision making of a physician can be analysed to get a deeper understanding of the process from initial data inspection to the treatment decision. Here, the think-aloud method can be used to assess the influence of a new visualization technique on the interventional strategy. Our paper focuses on quantitative evaluation on the example of aneurysm surface visualization, i.e. a measurable comparison of different visualization techniques.

The next step for the conduction of the quantitative evaluation is to check whether all requirements are met for inferential analysis (Figure 5, Stage 2). Examples for requirements are a clear hypothesis and a sufficient sample size [Fie09]. If these requirements are not fulfilled, descriptive statistics can be performed, comprising an analysis of the distribution of the data and an evaluation of measures for central tendency and variance. Appropriate visualizations for this information should be provided via box plots, bar charts and histograms. Even for inferential statistics, these visualizations should be presented to support the interpretation of the data.

In the following, the evaluation strategies for inferential statistics are explained in more detail, including problems in the medical fields and suggestions. Due to the wide variety of statistical tests with diverse assumptions about the data distribution, the sample size and the number of compared conditions, we only point out common tests and when to apply them. For a more detailed overview including a justification, we direct the interested reader to the book of Andy Field [Fie09], which includes further references for each test.

4.1. Parametric versus non-parametric tests

Parametric tests, such as a *t*-test, where differences between mean values are investigated, have more statistical power and, thus, a higher probability to reveal possible significances than non-parametric tests. However, they can only be applied if specific requirements are fulfilled, e.g. the sample size is sufficient and the data are scaled appropriately as well as normally distributed (Figure 5, Stage 3). In statistical practice, parametric tests are applied even if requirements are violated with the justification that these tests are robust against these violations [Fie09]. This makes it difficult for non-statistic professionals to decide when, e.g. a deviation from normal distribution is too strong and a sample size is too small, respectively. As a general suggestion, the *measure of central tendency* and the *scaling of the data* should be investigated. Different measures of central tendency comprise the mean, median and mode. A parametric test should only be considered if the *mean* is able to represent the central tendency. An example against this assumption is the usage of a forced-choice Likert scale (i.e. a neutral choice is missing) for data acquisition. Here, the mean value could lie between positive and negative ratings yielding the neutral choice that was prohibited in the initial setup. Thus, a misleading result would be reported. The median would be the appropriate measure of central tendency and a non-parametric test should be used. The scaling of the data can either be discrete (ordinal, nominal) or continuous (interval or ratio scale). For ordinal scaled data, such as ranked lists, the usage of a parametric test is debatable and, if in doubt, a non-parametric test is preferable. For continuous scaled data, a test of normal distribution accompanied by a visual inspection of the histogram should be performed [Fie09]. A possible test for this is the Shapiro–Wilk test, which examines whether the collected data came from a normally distributed population. Here, outliers should be considered as well. The additional visual inspection is necessary, since common small sample sizes in medical visualization rarely result in normally distributed data. Again, if the data significantly deviate from a normal distribution and the visual inspection is debatable, a non-parametric test is the preferable choice.

4.2. Independent, dependent and confounding variables

The controlled variation on the independent variable (also called factor) leads to changes to the dependent variable. In medical visualization, a typical independent variable is the visualization technique, whereas the different techniques are the respective conditions (Figure 5, Stage 4). The number of conditions affects the option to realize a *post hoc* test (see Section 4.6). A possibility for the controlled variation is the usage of an established visualization technique and a new one. This variation influences the dependent

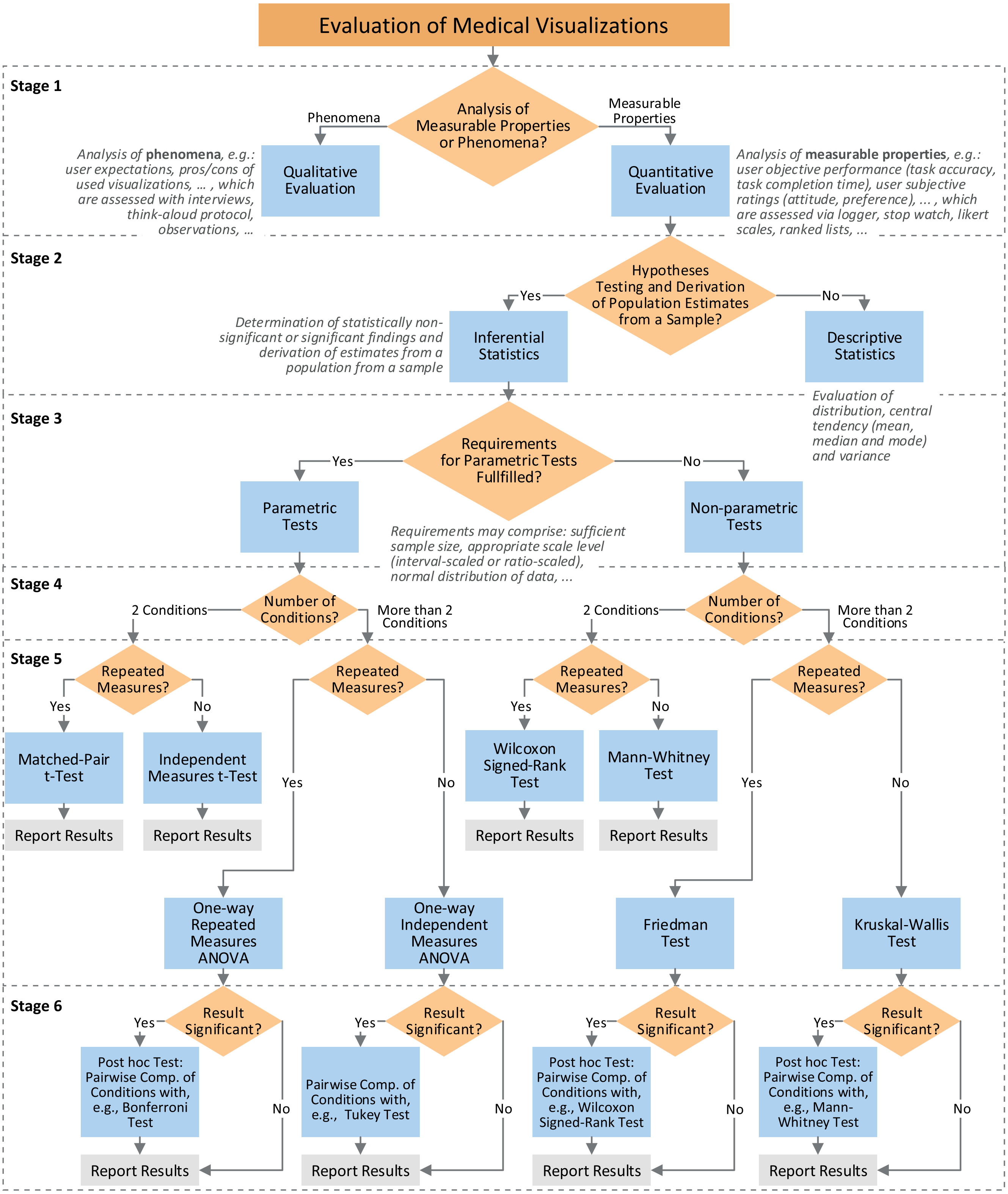
Evaluation of Medical Visualizations
Stage 1
Analysis of Measurable Properties or Phenomena?
Phenomena
Measurable Properties
Qualitative Evaluation
Quantitative Evaluation
Analysis of **phenomena**, e.g.: user expectations, pros/cons of used visualizations, ... , which are assessed with interviews, think-aloud protocol, observations, ...
Analysis of **measurable properties**, e.g.: user objective performance (task accuracy, task completion time), user subjective ratings (attitude, preference), ... , which are assessed via logger, stop watch, likert scales, ranked lists, ...
Stage 2
Hypotheses Testing and Derivation of Population Estimates from a Sample?
Yes
No
Inferential Statistics
Descriptive Statistics
Determination of statistically non-significant or significant findings and derivation of estimates from a population from a sample
Evaluation of distribution, central tendency (mean, median and mode) and variance
Stage 3
Requirements for Parametric Tests Fullfilled?
Yes
No
Parametric Tests
Non-parametric Tests
Requirements may comprise: sufficient sample size, appropriate scale level (interval-scaled or ratio-scaled), normal distribution of data, ...
Stage 4
Number of Conditions?
2 Conditions
More than 2 Conditions
Number of Conditions?
2 Conditions
More than 2 Conditions
Stage 5
Repeated Measures?
Yes
No
Matched-Pair t-Test
Independent Measures t-Test
Report Results
Report Results
Repeated Measures?
Yes
No
One-way Repeated Measures ANOVA
One-way Independent Measures ANOVA
Repeated Measures?
Yes
No
Wilcoxon Signed-Rank Test
Mann-Whitney Test
Report Results
Report Results
Repeated Measures?
Yes
No
Friedman Test
Kruskal-Wallis Test
Stage 6
Result Significant?
Yes
No
Post hoc Test: Pairwise Comp. of Conditions with, e.g., Bonferroni Test
Report Results
Result Significant?
Yes
No
Pairwise Comp. of Conditions with, e.g., Tukey Test
Report Results
Result Significant?
Yes
No
Post hoc Test: Pairwise Comp. of Conditions with, e.g., Wilcoxon Signed-Rank Test
Report Results
Result Significant?
Yes
No
Post hoc Test: Pairwise Comp. of Conditions with, e.g., Mann-Whitney Test
Report Results

Figure 5: *Guidelines represented as a decision tree with focus on quantitative evaluation (Stage 1). In Stage 2, the researcher decides if statistical significant findings are relevant or descriptive statistics are sufficient. The chosen statistical test depends on the collected data (Stage 3), the number of conditions (Stage 4) and the type of study (Stage 5). If more than two conditions were tested, post hoc tests are possible (Stage 6).*

variable, which can be measured. Examples for dependent variables are objective measures, such as required time, or subjective ratings, such as preferability. The possibilities of an unwanted influence to the dependent variable are called confounding variables. General examples are participant's motivation and study duration, which influence the performance [CW11]. Important confounding variables in medical visualizations are differences in the perception of participants, e.g. colour blindness, or differences on the used output device, e.g. display brightness, size and contrast. Usually, different domain experts, i.e. highly specialized physicians, are asked to participate in the study. Their varying experience also influences the measurement. A general method to reduce the influence of confounding variables is to keep them as constant as possible. This is easier for some than for others. Display types, lighting situation and an overall equal setting can be held constant in usability labs. Differences in experience can be controlled by using questionnaires, which try to quantize the experience to a certain degree. Then, it is possible to restrict the study to participants with similar experience. Another problem arises if a new visualization technique is compared to an established one, which could lead to a novelty bias against the new visualization. Here, thorough training sessions can reduce the bias. Ideally, they are carried out until the learning curve reaches a plateau. In summary, the approach of keeping confounding variables constant does not eliminate them, but exposes every participant equally to them. Thus, variations in the results, e.g. regarding accuracy, are theoretically explained by the studied factors alone. However, by controlling every aspect of an experiment, the external validity is reduced, i.e. how good the results are transferable to clinical practice. Here, researchers have to find the right balance between control and realism or perform several studies with different degrees of external validity.

4.3. Tasks and data sets

An evaluation task should represent main challenges of typical tasks as realistic as possible [CW11]. For medical visualizations, this assumption strongly limits the number of possible participants. A task imitating a real clinical scenario would require the know-how a physician gained during his education, training and experience. This aggravates in case of a special medical field, e.g. cerebral vessel pathologies. Here, an even smaller number of specialized surgeons and radiologists could participate. As a result, statistical analyses would lose power due to the small sample size. Therefore, the task is often approximated such that non-expert users can provide valuable test results. Typical examples for tasks related to medical visualizations are the estimation of size of pathologic structures for diagnosis or perceptually motivated tasks such as depth ordering of complex medical structures for intervention planning [PBC*16]. However, this limits the relevance and possibility for generalization [Bae15].

In conclusion, multiple similar tasks should be implemented to strengthen the result's plausibility and to enhance the external validity and reliability. For example, different aneurysms can be shown to evaluate a single aneurysm visualization technique. Here, particular care should be taken to create tasks with similar difficulty. Otherwise, this can be the reason for a higher variance in the results. Also, the aggregation of this acquired data should be analysed either run- or participant-related, which is explained in Section 4.5.

4.4. Experimental design

The type of experimental design can be divided into repeated measures design (within-subject), aiming at the variability of a particular value for the same individuals under different conditions, or the independent measures design (between-subject), aiming at differences between groups (Figure 5, Stage 5).

The choice of experimental design depends on the available participants and the evaluation goal. Independent measures studies avoid learning effects and the evaluation time is reduced for each participant compared to repeated measures design. However, groups of similar participants (w.r.t. age, experience, knowledge, etc.) have to be recruited. In the medical domain, these prerequisites are not easily met. Between-subject studies may suffer from interpersonal differences. Within-subject studies avoid these differences. Since they may suffer from learning or sequence effects, special care must be taken for the definition of tasks (e.g. the order of conditions across participants should be balanced) [CW11]. Although repeated measures designs are influenced by intra-personal differences (e.g. getting tired during the experiment), they may be superior to between-subject studies. When the same participants are involved and repeated measures are acquired, the overall variance is reduced and, thus, statistical significance can be reached more easily [Fie09].

In conclusion, repeated measures studies are recommended for the evaluation of medical visualizations due to the reduced variance and a lower number of required participants. However, certain evaluation goals such as the impact of surgical techniques on patients are not possible with repeated measures, since this surgery could only be carried out once for a single patient. An independent measures design should also be used if the risk of strong learning effects is too high. In medical visualization, this occurs if only a few data sets are available, which should be visualized with different techniques. Here, participants are able to recognize the data set and answer according to previous knowledge. Furthermore, an independent measures design is mandatory if the conditions are exclusive properties of the participants, e.g. physicians are either experts or novices. Differences regarding these groups can only be analysed if they are considered independently.

The chosen design ultimately influences the necessary statistical test that should be used to reveal differences between conditions. For example, acquired data that fulfill the requirements for a parametric test with more than two conditions and a within-subject design need to be analysed with a repeated measures ANOVA (an analysis of variance). An overview of the different test possibilities can be found in Figure 5.

4.5. Data aggregation choices

The acquired data of the study can be related to *participants* and to *runs* of a study, respectively. Data sets should be related to participants if the impact of the studied factors (e.g. different visualizations) on participants is investigated. In contrast, if general features of a technical system are evaluated, the results are independent of the participants and, thus, the data sets should be related to single runs. Depending on this distinction, the data should be aggregated or not.

For example, a medical visualization technique should be evaluated. To improve the reliability of the measured results, five runs with different medical data sets are performed. After presenting all data sets, five results are obtained. A common mistake is to handle these five results independently. However, since this evaluation scenario is participant-related, the results must be aggregated to a single value for each participant. Inappropriate data aggregation might bias the results of statistical tests. If ignoring the aggregation of participant-related evaluations, the sample size is artificially enlarged. This leads to an underestimation of the true data variance. Both artificial enlargement of the sample size and the underestimation of variance reasoned by the lack of data aggregation make statistical testing considerably more liberal, i.e. statistically significant results are obtained although no true effects exist [LSM16].

4.6. Using post hoc tests

Post hoc tests can be used optionally and are only possible if more than two conditions exist (Figure 5, Stage 6). More precisely, two conditions can be directly compared with each other (recall the tests contained in Figure 5, Stage 5). For more than two conditions, a first test reveals whether a statistically significant difference exists amongst them. Next, a pairwise comparison is carried out to compare the conditions against each other. For example, a repeated measures ANOVA for three conditions might reveal a significant difference between the conditions. If the researcher wants to identify which condition performed best or worst, the Bonferroni post hoc test is an appropriate method. This test compares pairwise mean values between each two groups with *t*-tests. A wide variety of post hoc tests exists [18 in SPSS 22.0 (IBM, New York, NY, USA)], making the right choice difficult. Figure 5 provides an overview of common statistical tests for this purpose. For more details, readers are referred to the book of Andy Field [Fie09].

5. Evaluation of 3D Aneurysm Surface Visualization

In the following, our exemplary quantitative user study is presented. We apply our guidelines described in the previous section.

5.1. Participants

The participants were recruited from visitors of the *Long Night of Sciences* in Magdeburg, Germany. During this event, scientific institutes present their research 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 five female and 29 male participants, with an age ranging from 16 to 66 years.

5.2. Independent and dependent variables

For our application, the independent variable is the aneurysm surface visualization with the three conditions Vis_A, Vis_B and Vis_C described in Section 3. The influence of experience with medical visualizations is used as a second independent variable. Here, we differentiate the medical visualization experience into the two conditions *MedVisExp* and *NoMedVisExp*. The two dependent variables comprising user objective performance are *required task completion time* and *accuracy*. The required time is logged after each completion of a task. We instructed our participants to take the time they needed. 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. As user subjective ratings, we used *suitability* and *preferability*. The ratings were assessed with a 5-point Likert scale ranging from $--$ (i.e. not suitable/preferable at all) to $++$ (i.e. very suitable/preferable).

5.3. Technical setup

The 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, 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 TabView GUI as well. Selection of visualization techniques and data sets for the participants was automatically carried out via Python scripts. The logging of participant's inputs and time required for each task were stored as text files.

5.4. Procedure and tasks

The GUI was presented to each participant, starting with a slide for the medical background information. Afterwards, examples of the three different visualizations Vis_A, Vis_B and Vis_C were shown. Each of the visualizations as well as the interaction, e.g. zooming and rotating, were explained in detail by the supervisor. The participants were 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 participant. Each participant had to solve 18 questions q_1–q_{18}, i.e. six per visualization, and had to decide which aneurysm possesses the larger volume. Finally, the participants answered a questionnaire comprising demographics questions and user subjective ratings.

5.5. Experimental design

For the comparison of the 3D visualizations, we use a repeated measures design. Here, each experiment is carried out such that all participants are confronted with each visualization technique six times. Thus, the amount of different visualization techniques shown is balanced. As a result, we repeat the question whether A_{Ref} is larger than A_{Comp} 18 times, which enhances the external validity. To reduce the influence of 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 provided in detail in our previous work [GSB*16]. 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 $P_1 - P_5$ as well as the order of segmentations were alternated. The pseudo-randomization ensures that each participant evaluates different data sets with varying segmentations, i.e. the participant 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 (recall Section 5.4), where the combinations of patient data and visualization techniques were not identical to the ones used in the test.

For the comparison regarding the medical experience, an independent measures design is used. This is necessary, since a participant cannot belong two both groups at the same time.

6. Results

Since our evaluation is participant-related, we aggregate the results of single participants (recall Section 4.5). The participants' answers form the set of observations for Vis_A, Vis_B and Vis_C. We count for each participant how many times he or she correctly answered for each visualization yielding numbers from 0 to 6. We also collect the set of averaged required times t_A, t_B and t_C that each participant needed for Vis_A, Vis_B and Vis_C. For each investigated aspect, we formulate the null hypothesis (H_0) and alternative hypothesis (H_1) explicitly. In the following, we explain the evaluation process according to our guidelines presented in Figure 5. We carry out inferential statistics (Figure 5, Stage 2) for all dependent variables:

- Non-parametric versus parametric tests (Stage 3). We test if the samples fulfill the requirements for a parametric test (properly scaled and normally distributed).
- Analysis of number of conditions (Stage 4). Based on the number of conditions, an appropriate test is chosen.
- Experimental design (Stage 5). We carry out the statistical test depending on the experimental design over all conditions.
- Post hoc test (Stage 6). If the statistical test indicates significant differences amongst the conditions, we carry out a post hoc test. Each condition is compared pairwise to assess the highest and lowest performing condition.

All statistical tests were carried out with SPSS 22.0.

6.1. Accuracy

6.1.1. *Differences regarding visualization*

The first analysis determines whether there exists a significant difference between the three visualization techniques w.r.t. the amount of correct answers, which range from 0 to 6. Box plots for the accuracy for Vis_A, Vis_B and Vis_C are provided in Figure 6 (left).

Non-parametric versus parametric tests (Stage 3) 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. The Shapiro–Wilk test yields the following significance levels:

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

Since all visualizations differ significantly from a normal distribution ($p < 0.05$), we use the non-parametric test for comparison.

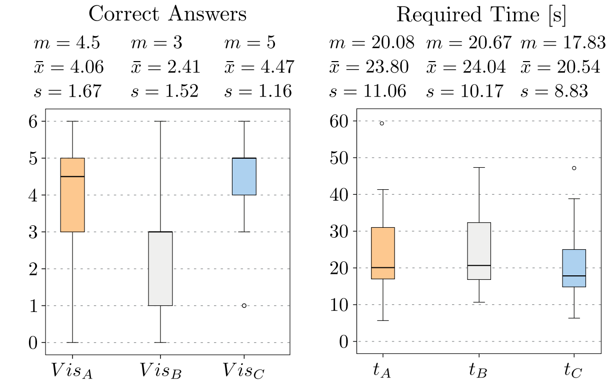

Figure 6: *Box plots of the accuracy (left) and the required time (right) for Vis_A, Vis_B and Vis_C including the median m, the mean $\bar{x}$ and the standard deviation s are shown.*

Analysis of number of conditions (Stage 4) The independent variable $visualization$ has the three conditions Vis_A, Vis_B and Vis_C. Therefore, tests for more than two conditions are considered.

Experimental design (Stage 5) According to our guidelines, we use the Friedman test, which compares the conditions based on ranks [non-parametric test, more than two conditions, repeated measures design (recall Figure 5)]. Here, we investigate if the visualization techniques lead to different results regarding accuracy. We define the hypotheses:

H_0: The participants achieve a similar accuracy with each visualization technique.
H_1: The participants achieve a different accuracy with the visualization techniques.

The Friedman test reveals that the accuracies significantly differ for the three visualizations ($\chi^2(2) = 25.38$, $p < 0.05$). Therefore, the hypothesis H_0 must be rejected.

Post hoc test (Stage 6) Since the visualizations lead to significant differences regarding accuracy, we compare each technique pairwise to identify the most suited. We use the Wilcoxon signed-rank test for Vis_A, Vis_B and Vis_C, which tests if their mean ranks differ. Because of the multiple tests, we use the Bonferroni correction method, i.e. adjusting the alpha by the number of comparisons (three comparisons yield one-third of 0.05 = .0167). The amount of correct answers is significantly higher for Vis_A ($m = 4.5$) than for Vis_B ($m = 3.0$) ($Z = -3.76$, $p < 0.0167$), where m denotes the median. Also, the amount of correct answers is significantly higher for Vis_C ($m = 5.0$) than for Vis_B ($m = 3.0$) ($Z = -4.07$, $p < 0.0167$). However, there is no significant difference between Vis_A ($m = 4.5$) and Vis_C ($m = 5.0$) ($Z = 0.95$, $p = 0.354$). Additionally considering the descriptive results, Vis_C ($\bar{x} = 4.47$, $s = 1.16$) performed better than Vis_A ($\bar{x} = 4.06$, $s = 1.67$).

Since Vis_B lead to the lowest results, we analysed how it competes with random guessing, where guessing would result in three correct answers. A Wilcoxon signed-rank test yields a significant difference ($Z = -2.09$, $p < 0.05$ with $\bar{x}_{Vis_B} < \bar{x}_{\text{guessing}}$). Thus, Vis_B

may systematically influence the participants to provide wrong answers.

6.1.2. *Differences regarding medical experience*

We want to investigate if there are significant differences regarding accuracy reasoned by experience with medical visualizations. The values of all three visualizations were averaged to a single value for each participant.

Non-parametric versus parametric tests (Stage 3) Every participant was assigned to one of the two experience groups. Only 10 of 34 participants had experience with medical visualization. Because of the small sample size in the experienced group, a non-parametric test is used.

Analysis of number of conditions (Stage 4) The independent variable *MedicalExperience* has the two conditions *MedVisExp* and *NoMedVisExp*. Therefore, tests for two conditions are considered.

Experimental design (Stage 5) Since each participant could be clearly matched to one of the experience groups, the measures were not repeated (between-subject). Thus, the Mann–Whitney test was used, which compares the sum of ranks of each group. We define the following hypotheses for experience with medical visualization:

H_0: The experience with medical visualization has no impact on accuracy.
H_1: The experience with medical visualization has an impact on accuracy.

The experience with medical visualization had no impact on accuracy ($Z = -0.99$, $p = 0.34$, *MedVisExp* $\bar{x} = 3.75$, $s = 0.83$; *NoMedVisExp* $\bar{x} = 3.40$, $s = 1.03$). Thus, we cannot reject H_0 and, thus, not accept the alternative hypothesis H_1. Since only two conditions were tested, no post hoc test and fifth stage is necessary.

6.2. Required time

6.2.1. *Differences regarding visualization*

We want to analyse whether there is a significant difference between the three visualization techniques w.r.t. the required time. Box plots for the required time for Vis_A, Vis_B and Vis_C are provided in Figure 6 (right).

Non-parametric versus parametric tests (Stage 3) 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,
- 0.007 for t_B and
- 0.006 for t_C.

All three variables significantly deviate from a normal distribution ($p < 0.05$). Therefore, we use a non-parametric test for comparison.

Analysis of number of conditions (Stage 4) Since the independent variable *visualization* has the three conditions Vis_A, Vis_B and Vis_C, tests for more than two conditions are considered.

Experimental design (Stage 5) For the analysis of accuracy regarding the visualization techniques, we use the Friedman test. The corresponding hypotheses are:

H_0: The visualization technique has no impact on the required time.
H_1: The visualization technique has an impact on the required time.

As a result, the Friedman test reveals no significant difference ($\chi^2(2) = 2.8$, $p > 0.05$). Thus, H_0 cannot be rejected. Since no statistically significant difference could be shown, we do not carry out a pairwise comparison of the required time. Comparing the descriptive data t_A, t_B and t_C, the participants performed the tasks on average faster with Vis_C ($\bar{x} = 20.54$, $s = 8.83$) compared to Vis_A ($\bar{x} = 23.80$, $s = 11.06$) and Vis_B ($\bar{x} = 24.04$, $s = 10.17$), respectively. Comparing the mean values of t_A and t_B, the participants required more time to fulfill the tasks with Vis_B.

6.2.2. *Differences regarding medical experience*

Similar to the accuracy, we want to investigate if there are significant differences regarding the required time reasoned by medical visualization experience.

Non-parametric versus parametric tests (Stage 3) and analysis of number of conditions (Stage 4) Both stages are identical to the one used for the accuracy (Section 6.1.2). Therefore, a non-parametric test for two conditions is used.

Experimental design (Stage 5) We define the following hypotheses:

H_0: The experience with medical visualization has no impact on the required time.
H_1: The experience with medical visualization has an impact on the required time.

Participants with experience in medical visualization performed the task faster ($\bar{x} = 20.67\,s$, $s = 7.23$) than participants without experience ($\bar{x} = 27.90\,s$, $s = 8.02$). This was reflected in a significant result of the Mann–Whitney test ($Z = -2.55$, $p < 0.05$) and the alternative hypothesis H_1 can be accepted. Since only two conditions were tested, no post hoc test is necessary.

6.3. Suitability and preferability

We want to investigate if there are significant differences in users' subjective ratings regarding our three visualization techniques. The collected data including the mode value, i.e. the answer ($--$, $-$, 0, $+$, $++$) that was given most often for each question as well as the amount of participants that provide answer $++$ and $+$ are shown in Figure 7.

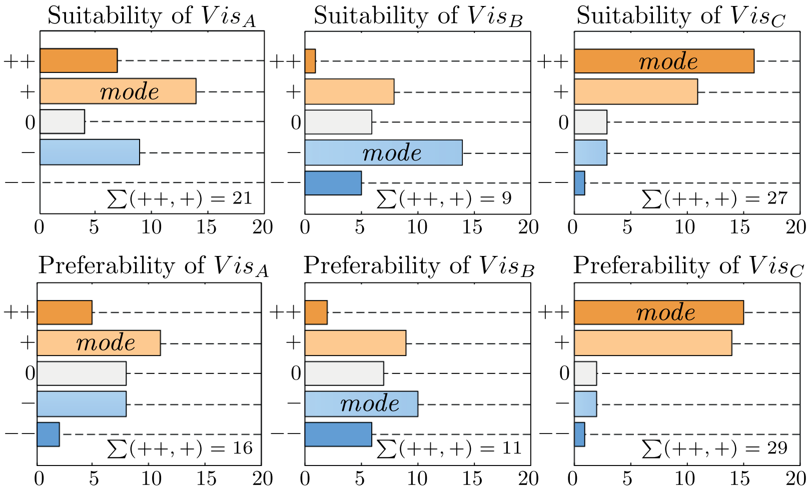

Figure 7: *Evaluation results of the participants regarding suitability and preferability of Vis_A, Vis_B and Vis_C. The mode value, i.e. the answer that was given most often for each question, is marked. Furthermore, the sum of answers ++ and + is provided.*

Non-parametric versus parametric tests (Stage 3) Since the users' subjective ratings were taken with a Likert scale representing an ordinal scale, a non-parametric test is used.

Analysis of number of conditions (Stage 4) Since the independent variable $visualization$ has the three conditions Vis_A, Vis_B and Vis_C, tests for more than two conditions are considered.

Experimental design (Stage 5) Given three conditions and repeated measures, a Friedman test is used. We define the following hypotheses:

H_0: Participants perceive the visualizations equally suitable.
H_1: Participants perceive the visualizations differently suitable.
H_0: Participants like the visualizations to similar extent.
H_1: Participants like the visualizations to different extent.

Participants mostly rated Vis_C with ++ for suitability and preferability, Vis_A with + for suitability and preferability as well as Vis_B with − for suitability and preferability. The amount of participants rating Vis_C as suitable and very suitable (i.e. answers are + or ++) was highest with 27, followed by 21 for Vis_A and nine for Vis_B. Similarly, the amount of participants 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. These differences were reflected in a significant result for both suitability ($\chi^2(2) = 21.76$, $p < 0.05$) and likeability ($\chi^2(2) = 18.37$, $p < 0.05$). Thus, we accept both alternative hypotheses.

Post hoc test (Stage 6) Next, we compare the visualization techniques to identify the most suitable and the most preferable one. We apply the non-parametric Wilcoxon signed-rank test. Reasoned by multiple testing, we use the Bonferroni-adjusted alpha (one-third of 0.05 = 0.0167). Participants perceived Vis_C significantly more suitable than Vis_B ($Z = -3.94$, $p < 0.0167$) and Vis_A significantly more suitable than Vis_B ($Z = -2.68$, $p < 0.0167$). Vis_C and Vis_A do not differ in terms of suitability ($Z = -1.86$, $p > 0.0167$). Although participants consider Vis_C and Vis_A equally suitable for size comparison of aneurysms, they like Vis_C significantly more than Vis_A ($Z = -2.80$, $p < 0.0167$). Moreover, participants liked Vis_C significantly more than Vis_B ($Z = -3.66$, $p < 0.0167$). No differences in terms of likeability could be found between Vis_A and Vis_B ($Z = -1.84$, $p > .0167$).

7. Discussion

The quantitative statistical analysis revealed significant differences of Vis_A, Vis_B and Vis_C w.r.t. accuracy, suitability and likeability. The pairwise comparison identifies that Vis_B performed worst regarding accuracy and suitability. For the required time, no significant differences were revealed. An explanation for this is that the participants were instructed to take as long as they need to choose the larger aneurysm. Although Vis_A and Vis_C were better than Vis_B and achieved similarly good results regarding these aspects, the participants liked Vis_C significantly more. Considering the central tendency measures alone, Vis_C is superior concerning all aspects and is therefore the best visualization technique for comparing two aneurysm surfaces. A possible conclusion might be that a derived quantity, i.e. the distance, improves the identification of the larger aneurysm. Additionally, Vis_C is the only visualization combining both surfaces into one. This may reduce the mental workload and supports perception of differences at the cost of information loss. However, the results indicate that this loss is acceptable.

Our analysis regarding medical visualization experience showed interesting results. Although no statistically significant differences could be identified with or without experience regarding accuracy, participants with experience performed tasks significantly faster. This indicates that participants benefit from prior knowledge.

Remarkably, Vis_B achieved a lower success rate than guessing. We assume that the participants did not understand the design of Vis_B. They might wrongly interpret the ghosting view and did not focus on the border areas but instead on areas facing towards them.

These areas are pre-dominantly colour-coded in cyan, since the A_{Comp} aneurysm is always drawn after the orange A_{Ref} aneurysm. Hence, Vis_B is inappropriate for comparison of aneurysm surface volumes.

8. Conclusion

The ultimate goal of medical visualization is the application in clinical practice to support diagnosis, treatment planning and fulfill information needs. Beneath qualitative evaluation, which is primarily applied in visualization [IIC*13], it is necessary to quantify the improvement of a new visualization technique with measurable and comparable properties, especially in accordance with the clinical approval procedure. In consequence, researchers should aim at quantitative evaluations whenever possible. In contrast, usually a small amount of physicians can participate in a study specialized in a sophisticated medical application. To overcome this limitation, the tasks of the user study should be simplified such that they are feasible for a broader range of participants and, thus, a quantitative evaluation. However, this happens at a loss of practical authenticity.

Our proposed guidelines allow for the comparative evaluation of three visualization techniques for the specific application of cerebral aneurysm volume assessment. 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 colour-coded surface visualization with quantitative distance information, such as our new technique Vis_C, supports the users in detecting the largest volume. This was reflected by a statistically significantly higher accuracy and better subjective ratings.

For future work, different approaches can be pursued. The visualizations can be improved, for example by including depth cues such as ambient occlusion. Furthermore, a systematic analysis of the influence of the aneurysm volume difference could identify whether a visualization may be well-suited for the depiction of large volume differences, but rather improperly suited for small differences. Finally, we are interested in a more comprehensive analysis on the influence of medical experience. Thus, a more differentiated acquisition should allow for investigation of a possible dependency regarding accuracy and required time. In the bigger picture, a discussion of effect sizes for each result would provide the strength of a significant result and, thus, benefit the comparison of evaluation results across different user studies.

Acknowledgement

We warmly thank Dr. Alexandra Baer for fruitful discussions and guidance for the evaluation. We thank Nico Merten for supporting us conducting the user study. 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

[Bae15] Baer A.: *Perception Guided Evaluation of 3D Medical Visualizations*. PhD thesis, University of Magdeburg, 2015.

[BBF*11] Busking S., Botha C., Ferrarini L., Milles J., Post F. H.: Image-based rendering of intersecting surfaces for dynamic comparative visualization. *The Visual Computer 27*, 5 (2011), 347–363.

[BCFW08] Bartz D., Cunningham D., Fischer J., Wallraven C.: The role of perception for computer graphics. *Eurographics (STARs)* (2008), 59–80.

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

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

[BRB*15] Berg P., Roloff C., Beuing O., Voss S., Sugiyama S., Aristokleous N., 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 (2015), 121008/1–13.

[BSV*17] Berg P., Saalfeld S., Voss S., Redel T., Preim B., Janiga G., Beuing O.: Does the DSA reconstruction kernel affect hemodynamic predictions in intracranial aneurysms? An analysis of geometry and blood flow variations. *Journal of NeuroInterventional Surgery* (2017), in print.

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

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

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

[CW11] Cunningham D., Wallraven C.: *Experimental Design: From User Studies to Psychophysics*. A. K. Peters, Ltd., Natick, MA, 2011.

[DRN*15] Díaz J., Ropinski T., Navazo I., Gobbetti E., Vázquez P.-P.: An experimental study on the effects of shading in 3d perception of volumetric models. *The Visual Computer 33*, 1 (2017), 47–61.

[Fie09] Field A.: *Discovering Statistics using SPSS*. Thousand Oaks, California: Sage Publications, 2009.

[GBNP15] Glasser S., Berg P., Neugebauer M., Preim B.: Reconstruction of 3d surface meshes for blood flow simulations of intracranial aneurysms. In *Proceedings of Computer and Robotic Assisted Surgery* (Bremen, Germany, 2015), pp. 163–168.

[GLH*14] Glasser S., Lawonn K., Hoffmann T., Skalej M., Preim B.: Combined visualization of wall thickness and wall shear stress for the evaluation of aneurysms. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2506–2515.

[GLR*09] Geers, A. J., Larrabide I., Radaelli A., Bogunovic H., Van Andel H., Majoie C., Frangi A. F.: Reproducibility of image-based computational hemodynamics in intracranial aneurysms: comparison of CTA and 3DRA. In *Proceedings of IEEE Symposium on Biomedical Imaging: From Nano to Macro* (Boston, USA, 2009), pp. 610–613.

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

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

[GSB*16] Glasser S., Saalfeld P., Berg P., Merten N., Preim B.: How to evaluate medical visualizations on the example of 3D aneurysm surfaces. In *Eurographics Workshop on Visual Computing for Biology and Medicine* (Bergen, Norway, 2016), pp. 153–162.

[GSK*15] Geurts A., Sakas G., Kuijper A., Becker M., von Landesberger T.: 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*, Vincent G. Duffy (Ed.), Cham: Springer (2015), pp. 336–344.

[IIC*13] Isenberg T., Isenberg P., Chen J., Sedlmair M., Moller T.: A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2818–2827.

[KHI*03] Kosara R., Healey C. G., Interrante V., Laidlaw D. H., Ware C.: User studies: why, how, and when? *IEEE Computer Graphics and Applications 23*, 4 (2003), 20–25.

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

[LBI*12] Lam H., Bertini E., Isenberg P., Plaisant C., Carpendale S.: Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics 18*, 9 (2012), 1520–1536.

[LEBB09] Lall R. R., Eddleman C. S., Bendok B. R., Batjer H. H.: 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 (2009), E2.

[LRHea05] Lang H., Radtke A., Hindennach M., Schroeder T., Fruhauf, N. R., Malago M., Bourquain H., Peitgen H. O., Oldhafer K. J., Broelsch C. E.: Impact of virtual tumor resection and computer-assisted risk analysis on operation planning and intraoperative strategy in major hepatic resection. *Archives of Surgery 140*, 7 (2005), 629–638.

[LSM16] Luz M., Strauss G., Manzey D.: Impact of image-guided surgery on surgeons' performance: A literature review. *International Journal of Human Factors and Ergonomics 4*, 3–4 (2016), 229–263.

[MMNG15] Miao H., Mistelbauer G., Našel C., Gröller M. E.: CoWRadar: Visual quantification of the circle of willis in stroke patients. In *Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine* (Chester, UK, 2015), pp. 1–10.

[PBC*16] Preim B., Baer A., Cunningham D., Isenberg T., Ropinski T.: A survey of perceptually motivated 3D visualization of medical image data. *Computer Graphics Forum 35*, 3 (2016), 501–525.

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

[PO08] Preim B., Oeltze S.: 3D visualization of vasculature: An overview. In *Visualization in Medicine and Life Sciences*, L. Linsen, H. Hagen and B. Hamann (Eds.), Berlin, Heidelberg: Springer (2008), pp. 39–59.

[SOBP07] Schumann C., Oeltze S., Bade R., Preim B.: Model-free surface visualization of vascular trees. In *Proceedings of Eurographics Symposium on Visualization* (CA, USA, 2007), pp. 283–290.

[syn16] syngo Application Software. Operator Manual, VD11. Siemens Healthcare GmbH, 2016.

[WT05] Weigle C., Taylor R. M.: Visualizing intersecting surfaces with nested-surface techniques. In *Proceedings of IEEE Visualization* (Minneapolis, MN, USA, 2005), pp. 503–510.

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