A Critical Analysis of the Evaluation Practice in Medical Visualization

Bernhard Preim, Timo Ropinski, Petra Isenberg

Eurographics Workshop on Visual Computing for Biology and Medicine (2018), pp. 1–12
A. Puig Puig, T. Schultz, and A. Vilanova (Editors)

A Critical Analysis of the Evaluation Practice in Medical Visualization

B. Preim [1] and T. Ropinski[2] and P. Isenberg[3]

[1]University of Magdeburg & Research Campus STIMULATE, Germany
[2]University of Ulm, Visual Computing Group, Germany
[3]Inria, Aviz research group, Saclay, France

Abstract

Medical visualization aims at directly supporting physicians in diagnosis and treatment planning, students and residents in medical education, and medical physicists as well as other medical researchers in answering specific research questions. For assessing whether single medical visualization techniques or entire medical visualization systems are useful in this respect, empirical evaluations involving participants from the target user group are indispensable. The human computer interaction field developed a wide range of evaluation instruments, and the information visualization community more recently adapted and refined these instruments for evaluating (information) visualization systems. However, often medical visualization lacks behind and should pay more attention to evaluation, in particular to evaluations in realistic settings that may assess how visualization techniques contribute to cognitive activities, such as deciding about a surgical strategy or other complex treatment decisions. In this vein, evaluations that are performed over a longer period are promising to study, in order to investigate how techniques are adapted. In this paper, we discuss the evaluation practice in medical visualization based on selected examples and contrast these evaluations with the broad range of existing empirical evaluation techniques. We would like to emphasize that this paper does not serve as a general call for evaluation in medical visualization, but argues that the individual situation must be assessed and that evaluations when they are carried out should be done more carefully.

Categories and Subject Descriptors (according to ACM CCS): Empirical Evaluation

1. Introduction

As medical visualization has matured, evaluations have become increasingly important. In particular, the large variety of medical visualization techniques available, requires a better understanding how these techniques compare to each other and how they provide value in practical use. Assessing tools' complexity in regular use and identifying fit with end users' information needs and workflows would help to tease apart overengineered solutions from those with practical value for end users. Given the few in-depth evaluations, industrial software engineers must often rely on their intuition when selecting published medical visualization techniques to be used in their products or conduct their own evaluations.

Thus, within this paper we provide a critical discussion of the state of empirical evaluations in the field of medical visualization. *Empirical evaluation* covers all evaluation concepts where target users or related stakeholders are in the focus. The HCI community has a long tradition of empirical evaluation and a number of in-depth investigations related to the planning, conduct and (statistical) analysis of empirical evaluations are available [LFH17, Shn10]. The visualization community (VIS) can build on this experience. However, evaluating visualization techniques and whole visualization systems has special aspects that make the empirical evaluation challenging [Pla04]. To address these challenges, the BELIV (Beyond time and error: novel evaluation methods for visualization) workshop series was initiated in 2006 (an ongoing biennial workshop), with the main goal to improve empirical evaluation and to increase the diversity of the applied methods. As an example of a specific method, insight-based evaluation is employed to analyze the knowledge discovery capabilities of information visualization systems [Nor06]. Also various other techniques have been adapted to be specific for VIS, including interaction log analysis, eye tracking and long-term case studies.

In general, while the evaluation-related research in visualization is still largely focused on information visualization [AS04, Car08, LBI*12, Mun09] as well as visual analytics, few research is carried out to analyze and advance the practice of evaluation in scientific visualization techniques, e.g., the interactive handling of spatial data, such as volume, flow and tensor data. A notable exception is the survey by Isenberg et al. which explicitly includes visualization techniques focusing on spatial data [IIC*13]. They have analyzed existing publications w.r.t. the conducted evaluation, and among these considered two medical visualization papers as good examples: the discussion of the medical visualization table [LRF*11] and super-

quadric tensor glyphs [SK10]. While we will discuss several medical visualization papers with an empirical evaluation, we explicitly do not consider our paper as a survey, but rather as a discussion of few illustrative examples. The papers we consider here are based on medical image data, i.e., we do not discuss visual analytics in health care, e.g., based on electronic health records or prescription behavior.

In this paper, we give an overview of empirical evaluation methods as conducted in the area of information visualization (Sect. 2), and discuss how these methods can be applied to medical visualization. We discuss how single medical visualizations and entire medical visualization systems are currently evaluated (Sect. 3). In Section 4 we discuss selected papers with a substantial evaluation in more detail. The comparison of empirical evaluations in information visualization and medical visualization leads to thoughts for a research agenda for medical visualization (Sect. 5).

2. Empirical Evaluation in Information Visualization

In the area of information visualizations researchers have intensively discussed the evaluation topic from several angles: based on methods and methodologies [And08, Car08, TM04], on evaluation goals [LBI*12], or when in the development cycle to conduct evaluation [Mun09]. Andrews describes the space of empirical evaluations based on the following dimensions:

- at which stage in the development process it occurs,
- the character of the evaluation (qualitative, quantitative),
- the duration of the evaluation, and
- the context of the evaluation.

In the following subsections, we discuss these dimensions.

2.1. Stage

Besides Andrews, also Munzner [Mun09] analyzed evaluations with respect to their occurrence in the development cycle. Based on her own experience and a review of existing evaluation practices, she proposes a *four layer nested model*, where different kinds of evaluation are performed within the four levels of visualization design. Her levels roughly correspond to stages, but consider iterative cycles where the design at the same level may be repeated. The evaluation methods are tightly coupled and appropriate for a particular level, such as domain characterization, data abstraction, selection of a visual representation and the specific algorithm. Andrews in contrast only differentiates two phases, and thus defines formative and summative evaluations.

Formative evaluation. When evaluations are part of an ongoing development process and designed to identify strong and weak aspects of the current prototype (sketches, storyboards, click prototypes or running systems), they are referred to as *formative evaluations* [Lew12]. The major goal is to identify how a single visualization or a system can be improved. Open or semi-structured interviews, possibly enhanced with video or think-aloud recordings, are major instruments in a formative evaluation. Users may be explicitly asked for redesign suggestions, which may steer the further development in a constructive manner. A convincing evaluation strategy would be to conduct a formative evaluation, and plan additional time to improve the prototype according to major problems identified. In medical visualization research papers we rarely see this type of evaluation.

Summative evaluation. When an evaluation is carried out at the end of a project to enable an assessment of the usefulness and user experience of a working system, it is referred to as *summative evaluation* [Lew12]. Rather few users are required to reliably identify most usability problems [HS10]. Summative evaluations are often performed as controlled lab studies with standardized instructions, conduction and (often statistical) data analysis; or as expert reviews.

Both summative and formative evaluations have been carried out in medical visualization. However, both types of evaluations are often conducted perfunctorily. Few medical visualization evaluations, for example, have been carried out with a large number of participants actually representing the target user group, namely specialized physicians. Instead, physicians are often "substituted" by medical students or — even more convenient and potentially misleading — by computer science students. Often, also web-based questionnaires are used where the authors have no control over the selection of participants and other important details, such as monitor type, resolution and lighting, e.g., in an evaluation from Ritter et al. who analyzed visualization techniques for vascular structures [RHD*06]. Despite the large number of participants (160) in this past study, the *external validity*, that is the transfer to real problems and working contexts remains unclear due to a potential sampling bias. More generally, it must be worried whether highly complex visualization techniques can be adopted by the true target user group or will end up serving more as examples of research and engineering capability than for practical use. Statistically significant results can be misleading when the study design and execution are not asking the right questions or sample the wrong type of test subjects.

A problem, in particular in summative evaluation of medical visualization systems, is the requirement to develop a complete system which captures the entire workflow from data import over analysis to result documentation as well as the availability of experts and their willingness to cooperate in the use of a new system. Naturally, such system development leads to a large overhead as compared to evaluations that target isolated visualization techniques. However, tradeoffs are possible; isolated visualization techniques along with interaction facilities to adjust their parameters can be made available to physicians.

2.2. Evaluation Character

There are two fundamentally different characters of evaluations: qualitative and quantitative evaluations.

Qualitative evaluation. When evaluations aim at obtaining insights on how new techniques are adopted, which problems are solved with them, and whether users are satisfied with them, they are often of qualitative nature. Observation is one of the most fundamental techniques of qualitative work. Asking users to think-aloud and thus to comment what they are doing including the underlying motivation is one potentially useful method to collect data during observation. Video and audio recordings can also be enriched by screen recordings and log files. Next, researchers have to code, classify, and align

the collected data, which is a particularly challenging aspect of such evaluations. Thus, qualitative studies are typically very involved studies that collect a huge amount of heterogeneous data and need very careful and tailored analysis (a statistical analysis in comparison has very little data). To assess medical visualizations, qualitative methods may be used to study a wide range of aspects, e.g.

- whether users trust in a particular medical visualization,
- how a tool helps in medical decision making, or
- how they assess the utility of a particular medical visualization.

While qualitative methods in computer science were for a long time barely considered as a valid scientific method, in the meantime they are appreciated, especially in the HCI community, based on the holistic insight that they may deliver. In the visualization domain, however, qualitative studies are often considered as not rigorous or difficult to carry out — whereas the opposite is true.

Eye tracking is another instrument that is potentially useful to analyze viewing patterns. It is widely used in information visualization, see e.g. the survey by Blaschek et al. [BKR*14], but also in medical imaging research where it is used, for example, to study viewing patterns of younger and more experienced radiologists related to breast cancer or lung nodules [Kru96]. Blaschek et al. mention 109 references from many visualization subfields, such as graph visualization, geographic visualization, but also dynamic and 3D visualization where eye tracking was used. Neither in this survey nor anywhere else we could find a medical visualization paper that employs eye tracking. While in earlier times eye tracking was subject to drift and other sources of significant inaccuracy it made considerable progress and clearly would be a viable instrument to assess the perception of medical visualization.

Quantitative evaluation. When an evaluation primarily yields numbers, they are referred to as quantitative evaluations. Often task completion times and error rates are analyzed with frequent statistical approaches such as p-values or descriptive statistics, while likelihood estimation and Bayesian inference are beginning to gain importance. Even subjective opinions such as preferences can be turned into quantitative measurements through Likert scales and then be analyzed statistically.

Quantitative and quantitative evaluation in medical visualization. We already mentioned the problem of selection bias (among the participants of a study) above. Other sources of bias inherent in quantitative experiments relate to the tasks and data that are used. Since the developers of a new technique decide on this selection, they may even unconsciously select datasets and tasks where their technique probably has an advantage over other techniques. As a consequence, it is often unlikely that an independent research group can accurately reproduce the results. In medical visualization quantitative evaluation dominates, which gave rise to a survey paper [PBC*16] and a guideline-type of paper [SLB*18] that explains the specific selection of statistical methods for different scenarios. The focus of this work are perception-based evaluations, i.e. depth and shape perception tasks where time completion and error rates are assessed. Most recently, Meuschke et al. [MSL*18] partially automated such experiments, in particular depth perception experiments related to vascular visualizations. As part of their system, pairs of points are generated that are employed for the depth perception tasks. These pairs represent simple, moderately difficult and difficult tasks. Qualitative evaluations in medical visualization typically provide short informal feedback where details often are missing, as to how questions were prepared. On the positive side, even these short qualitative evaluations can be directed towards realistic tasks.

2.3. Evaluation Duration

In a *short term* evaluation participants often perform perception-related tasks, that for example involve judging the encoding of depth in 3D images through an image comparison, e.g., in a depth judgment study. In a *long-term evaluation*, a system is observed and analyzed with methods derived from ethnographic field studies [NO99]. Medical visualization evaluations are typically short-term, i.e., participants get instructed, use a system supervised by an instructor for a short period of time and provide anecdotic feedback as well as measures, such as task completion times. The lack of long-term evaluations is a further severe shortcoming of evaluations in medical visualization.

In HCI, (long term) ethnographic studies have been conducted since more than two decades (see e.g. [WS96] who found interesting and un-intended patterns related to e-mail use). In information visualization, Shneiderman and Plaisant made a case for long-term case studies to evaluate visualization systems [SP06]. They discriminate long-term case studies that are carried out in a rather small time (four weeks) from studies that last up to a few years. Users document how they use a system and their experience in a *diary of use* and they are regularly interviewed (and reminded to perform the documentation). Such an evaluation style is appropriate for whole visualization systems, e.g. for the diagnosis of vascular diseases or virtual colonoscopy. While evaluations that last many months are outside the scope of a Phd study, a four weeks evaluation with weekly interviews should be feasible.

2.4. Evaluation Context

An evaluation may be performed in a lab with controlled conditions or in a more realistic context, e.g., in a tumor board discussion or a briefing room where surgeons prepare for the upcoming surgery. Since human thinking is strongly guided by associations (the environment may remind us of some aspects), this context plays an important role for cognitive processes and, thus, the more long-term qualitative evaluations discussed earlier. Also constraints, such as the sterile situation in an operating room, are essential, e.g., for a visualization intended to be used intraoperatively. For practical reasons, i.e. simplicity, evaluations are often carried out in a lab, but this clearly reduces the external validity.

3. An Overview of Evaluation in Medical Visualization

In the following, we give an overview of empirical evaluations or discuss the absence of such evaluations in medical visualization publications from the IEEE VIS and the EuroVis conference. We considered those papers where the medical application is stated in the title, based on terms such as “surgery”, “vascular” or “medical”. We excluded papers that aim at an improvement of surface or volume rendering without an explicit link to particular medical applications.

We chose to focus on IEEE VIS and EuroVis , since authors usually make a substantial effort to get their work published in these venues and acceptance rates are low. Accordingly, we consider that only evaluations of a high standard pass through the reviewing process — and that authors knowingly put effort in careful evaluation.

While in earlier years (before 2010) the development of new medical visualization techniques without any empirical evaluation was typical, in recent times, empirical evaluation is more often included in the (accepted) papers. In some of the earlier papers, e.g., [HMK*97, BWKG01, KFW*02], a user study is "hidden" in a result section where it is discussed along with many other things, such as the performance of algorithms. Today, it is more typical that a section is devoted to a "User Study", "Informal Evaluation" or "Case Study". The depth of these evaluations is typically not high (we mention positive examples in Sect. 4) and the evaluations are add-ons to a predominantly technical paper.

One type of empirical evaluation was used more often: task-based experiments to assess perceptual properties of medical visualization techniques, e.g., whether a new vessel visualization technique improves depth perception or shape perception [BGP*11, RSH06]. These evaluations are useful, but restricted to low-level perceptual aspects. Thus, these evaluations do only provide little hints whether a visualization technique is helpful in a clinical context, which involves further cognitive activities [AS04], such as diagnosis or treatment planning. One may argue that perception-based evaluations are a basis for choosing visualization techniques to be later integrated in systems that are evaluated more comprehensively. While this is possible in theory, there is no documented example where this actually happened. In contrast, entire visualization systems typically use much simpler techniques than the latest research results indicate. One reason for this situation may be that the task-based experiments are not only low level but rather unrealistic: visualization techniques are compared by means of *static images*, i.e. the essential depth cues related to rotating 3D visualizations are ignored. Since there is already a survey article on perception-based evaluations [PBC*16], we do not discuss them here in detail. To provide a structured analysis, we discriminate early and more recent papers. We systematically analyze the publications from the visualization field. Medical 3D visualizations are also evaluated in a few medical journal papers or in papers that appeared in the International Journal of Computer-Assisted Radiology and Surgery, a journal that presents interdisciplinary work between computer scientists and physicians. These publications discuss the value of interactive 3D visualizations for high level cognitive tasks, such as learning [AA16, PS18, SHM09] and surgery planning [HZP*14, LGF*00, PWN*15]. These evaluation strategies, however, are beyond the scope of this paper.

3.1. Evaluations in Early Medical Visualization Papers

In the following we briefly discuss early high-impact medical visualization papers from 1997 up to 2009 with a focus on evaluation. However, most of them do not contain an *empirical* evaluation, i.e., documented feedback from users.

[HMK*97]	Virtual Voyage: Navigation in the Human Colon
[BWKG01]	Non-linear colon unfolding
[KFW*02]	CPR: curved planar reformation
[KWFG03]	Advanced curved planar reformation: Flattening of vascular structures
[VBvdP04]	DTI visualization with streamsurfaces and evenly-spaced volume seeding
[HQK06]	A Pipeline for Computer Aided Polyp Detection
[RHD*06]	Real-Time Illustration of Vascular Structures
[LWP*06]	Full Body Virtual Autopsies using a State-of-the-art Volume Rendering Pipeline
[BHW*07]	Feature Emphasis and Contextual Cutaways for Multimodal Medical Visualization
[RRRP08]	Interactive Visualization of Multimodal Volume Data for Neurosurgical Tumor Treatment
[JSV*08]	Novel interaction techniques for neurosurgical planning and stereotactic navigation
[RHR*09]	Multimodal Vessel Visualization of Mouse Aorta PET/CT Scans

Hong et al. [HMK*97] provided a comparison of virtual endoscopy images with views from optical endoscopy of the same data. Thus, a major argument for the presented technique is that it creates visualizations that are as close as possible to their real-world counterpart to achieve a similar diagnostic accuracy compared to real endoscopy. Gastroenterologists and radiologists thus represent the target user group. Like the later paper on virtual endoscopy by Vilanova et al. [BWKG01], no empirical evaluation was carried out. The potential of the methods is just shown by applying them to representative data. Curved planar reformation [KFW*02] and its refinements [KWFG03] serve to diagnose vascular diseases more efficiently. Also these techniques were only analyzed w.r.t. the influence of noise in the underlying data and the visual appearance but without any feedback from the target user group. The visualization of Diffusion Tensor Images with streamlines and streamsurfaces [VBvdP04] was motivated by neurological diseases where the fiber tracts should be displayed in their spatial context. The considerable algorithmic challenges and a viable solution were discussed by Vilanova et al., but user feedback was not included. Hong et al. [HQK06] provide an analysis of the accuracy of polyp detection (more like an image analysis paper), but no empirical evaluation of the visualization and user interface.

Ritter et al. [RHD*06] as well as Joshi et al. [JSV*08] presented innovative vessel visualization techniques and showed that their new techniques improve depth perception in depth judgment tasks.

Ljung et al. [LWP*06] presented a rather new application area, namely the use of radiological image data and advanced visualization for forensics. A strength of the paper is the domain characterization that includes a discussion of causes of death after a crime, e.g. entry and exit wounds of a shooting, subtle fractures and other traces. The authors came up with a technically demanding solution that enables the interactive handling of the large full-body datasets. The examples used for the discussion of the method are carefully chosen and probably clinically relevant. As a further example for an influential medical visualization paper of this period, we mention the work by Burns et al. [BHW*07] who presented methods to emphasize features in medical volume data. They showed the influence of various parameters and generated impressive images. Neither in the paper nor follow-up publications was any evidence given that a physician or student of medicine actually used these vi-

sualizations. Rieder et al. [RRRP08] describe a system that employs multimodal medical image data and related fused 3D visualizations for brain tumor surgery. They include a short empirical evaluation and mention that their system was discussed with three physicians, but for none of the statements it is clear whether this represents the opinion of only one or more participants and whether the single opinion is from an expert user or a novice. To show the value of their vessel analysis system, Ropinski et al. [RHR*09] have exemplarily described three application cases. These application cases contain an in-depth discussion of the medical findings, which were enabled by the proposed system. While this demonstrates its usefulness for the collaborating medical partners, it does not show that the same system would also be valuable for a large and diverse user group.

In summary, early medical visualization papers aimed at solving algorithmic and other technical challenges, often based on a good domain characterization, but often without any user feedback, except for task-based perceptual evaluations. We discuss selected empirical evaluations, including those of early papers in Sect. 4.

3.2. Evaluations in Recent Medical Visualization Research

In this subsection, we discuss selected more recent medical visualization papers (since 2010), again with a focus on their evaluation component, and try to identify common strategies, differences and changes compared to early medical visualization papers. In general, evaluation gained importance and we found more medical visualization papers at IEEE VIS or EuroVis in that period that contain some sort of *empirical evaluation*. The depth and quality of the evaluations, however, is mixed. While the comparison with previous methods, often carried out as a visual comparison, plays a large role, substantial evaluations in realistic settings are still rare. In the following, we discuss the empirical evaluation of selected papers. Like in the previous subsection, the selection process also considers *diversity*, i.e., we do not only consider specific subfields or limit our analysis on the results of particular research groups, such that we achieve a representative sample.

[GNBP11]	FlowLens
[KGP*13]	Vortex Extraction in Cardiac Blood Flow Data
[MMV*13]	Vessel Visualization using Curvicircular Feature Aggregation
[SzBBKN14]	Focus-and-Context Visualization for 3D Ultrasound
[RvdHD*15]	Visual Analytics for the Exploration of Tumor Tissue Characterization

FlowLens. Gasteiger et al. [GNBP11] were motivated by the large number of potentially interesting attributes of blood flow: local differences in pressure, velocity, or in the residence time that characterizes how long the flow stays, e.g. in a pathologic region. They concluded that focus-and-context renderings are beneficial to understand one flow feature in the context of another and suggested the *FlowLens* to steer a lens region. The *FlowLens* is a semantic lens, i.e. in the lens region different attributes are presented compared to the surrounding. The design involves many decisions, such as how to display the lens and the selection of lens shapes. To verify such design decisions, informal interviews with two physicians and one biomedical researcher were carried out. They focused on lens design and transformation as well as on different scopes in which the lens could be applied. Even the two experts considered details, such as color scales and illumination, differently and, thus, options are needed to adjust the *FlowLens*. Since the users were asked to use the system themselves after an instruction it became obvious where they had difficulties.

Vortex Extraction in Cardiac Blood Flow Data. Köhler et al. [KGP*13] assessed vortex extraction methods to reliably extract vortices from measured cardiac blood flow data. Visualizations were generated that emphasize near-vortex regions motivated by clinical research that describes correlations between the location, extent and temporal behavior of vortical flow and pathologies. The "Informal evaluation" section mainly describes two cases in detail, including quite extensive information on the patient history and previous examinations that are clinically relevant for the diagnosis. This qualitative evaluation served to analyze how the visualizations were used to support disease understanding, how the symptoms relate to the extracted flow features and also which additional examinations may be necessary to answer questions arising from the flow visualization. This work was integrated in *BloodLine*, a system that is in clinical use since that time.

Vessel Visualization using Curvicircular Feature Aggregation. Mistelbauer et al. [MMV*13] introduced a method that summarizes curved planar reformation images in a single static image. This technique is advanced and potentially useful, since interaction is reduced and more efficient analysis is possible. A summative evaluation based on phantom and clinical data was performed where the new technique was compared to Maximum Intensity Projection and Curved Planar Reformation. Nine radiologists were involved in the user study that was based on a comprehensive questionaire (48 questions) which is available online which contributes to the reproducibilty of this evaluation. The short time for the evaluation does not allow definitive statements about the usefulness of the new method, since it clearly requires more learning effort.

Volume rendering ultrasound data. Schulte zu Bergen et al. [SzBBKN14] introduced volume rendering techniques for ultrasound data. Instead of a separate evaluation they discussed selected datasets in a result section and commented on aspects clinicians are interested in, e.g. "seeing the bone surface in context with the muscle" for a shoulder dataset, "the path of the carotid artery and its bifurcation in a spatial context" and "the achilles tendon in its whole shape to identify possible tears": They use static images to evaluate whether physicians can recognize the details important to them. Thus, the *recognizeability of details* may be a generally useful criterion.

Visual Analytics for the Exploration of Tumor Tissue Characterization. Raidou et al. [RvdHD*15] deal with tumor tissue characterization which is relevant for radiation treatment planning and thus both for physicians and medical physicists who are jointly responsible for a treatment plan. They derive a number of features relevant for the tumor tissue and employ dimension reduction to display them in 2D as part of a multiple view framework. Different risk zones should be displayed and the selection of color scales is an important issue. The evaluation was carried out with ten participants, including two research physicists and three medical physicists (the narrower target audience). The participants could use the tool on their own. Think-aloud was used to reveal the impression of the users. The participants not only comment on the evaluated system

but also relate it to tools that they employ for in-depth planning of radiation treatment. In addition, semi-structured interviews were carried out. Real clinical data, e.g. a prostate and a cervical cancer case, were used. The evaluation provides a good impression of the possible use of this tool for medical research.

4. Selected In-depth Empirical Evaluations

In this section, we describe further evaluations that contain interesting aspects. We discuss in particular, criteria that may be useful for (many) other evaluations of medical visualization systems. The selection of papers includes some early and some more recent publications. Also individual techniques, as well as whole systems are included. The following medical visualization papers contained a broader discussion of user feedback:

[HPSP01]	Visualization and Interaction Techniques for the Exploration of Vascular Structures
[OP05]	Visualization of vasculature with convolution surfaces: method, validation and evaluation
[BHWB07]	High-quality multimodal volume rendering for preoperative planning of neurosurgical interventions
[LLPY07]	Uncertainty Visualization in Medical Volume Rendering Using Probabilistic Animation
[KKPS08]	Sinus endoscopy: application of advanced GPU volume rendering for virtual endoscopy
[vPBB*10]	Exploration of 4D MRI Blood Flow Using Stylistic Visualization
[LRF*11]	MedVizTable
[HMP*12]	BiopsyPlanner
[SLK*17]	PelVis: Atlas-based Surgical Planning for Oncological Pelvic Surgery

Visualization and Interaction Techniques for the Exploration of Vascular Structures. Hahn et al. [HPSP01] approximate vascular trees with truncated cones, thus assuming a circular cross-section. This method was heavily used for liver surgery planning and the observation of surgeons using the technique revealed that they often carefully look at close-ups, in particular at bifurcations. Realistic cases (tumor surgery) were used for the evaluation. However, the approach of gathering feedback was very informal and not well-structured. The results of this *formative evaluation* lead to incremental changes, e.g. smoothing of the vessel radius and adding a cap at the terminal branches aiming at increasing the realism. In an "application section" four examples of the developed vessel visualization for liver surgery planning were discussed. The good characterization of anatomical variants and the recognition of the vessels' branching structures were identified by physicians as an advantage. From the text, however, it is not clear how many surgeons were involved, let alone any demographic characteristics (sex, age, experience).

Visualization of Vasculature with Convolution Surfaces. The development of this technique has its origins in the previously described work (recall [HPSP01]). Discontinuities at the connection of truncated cones cannot be avoided with an explicit construction of vascular trees. Since surgeons typically strongly zoomed in the visualizations to recognize details (they zoomed stronger than the developers expected), these discontinuities were actually considered disturbing. Thus, an implicit visualization technique was developed and refined w.r.t. unwanted blending effects and performance issues [OP05]. Accuracy and user feedback were discussed in some detail. 11 physicians (radiologists and surgeons) filled a questionnaire after looking at 10 series of (static) images where the new vessel visualization technique was compared with a state-of-the art method [HPSP01] and standard surface rendering. The following criteria were used:

- clarity,
- similarity to intraoperative views,
- comprehensibility of the spatial relations and
- visual quality

The distinction between "clarity" and "visual quality" is not sharp; probably these features are overlapping. The image series compared highly zoomed in close-up views in the surrounding of bifurcations. The same experiment with another zoom factor may result in strongly different results. Thus, the specific choice of stimuli needs to be carefully discussed. The major limitation of this evaluation was that just static images were compared. Thus, physicians have not actually used the visualizations for realistic tasks.

High-quality Multimodal Volume Rendering. Beyer et al. [BHWB07] discussed two cases in the "results" section, where a neurosurgeon used "his" multimodal data (CT and MRI data) and the presented tool to plan a surgery. Various screenshots from the planning stage and some comments of this user give an impression how the system may be used. The surgeon commented on how he defined the resection border and identified "no touch" areas that should be preserved in surgery. The cases are carefully chosen and include an easier and a more challenging case (deep-seated tumor). The paper also reports on the necessary setup time for image analysis and preparation of an initial visualization and comments on major drawbacks, namely that registered CT and MRI data are needed and that the adjustment of the visualizations involve many parameters. The surgeon commented on the high correlation to intraoperative views. Thus, whether images generated during planning resemble the intraoperative situation is again considered an essential criterion (recall [OP05]). However, it is not clear how similarity is actually assessed. Should transfer functions and color selection for the surface rendering be chosen such that the colors appear natural?

Uncertainty Visualization in Medical Volume Rendering Using Probabilistic Animation. Lundstroem et al. [LLPY07] address the problem that volume renderings of vascular structures are highly sensitive to details of the transfer function (TF) specification. Thus, with a predefined TF for displaying contrast-enhanced CT data, a stenosis or even an occlusion of the vessel may appear, whereas with a slightly changed TF no pathology is visible. Lundstroem et al. aim at a clinically feasible solution, i.e. a simple and efficient visualization of alternative renderings. They enable the user to specify a sensitivity lens and (only) inside the lens region an animation presents slightly different renderings with modified TF. Simulated and clinical data are used for the evaluation with 12 physicians (11 radiologists, one cardiologist). The users had the task to detect and localize stenosis as well as to assess their severity, since the degree of a stenosis determines the treatment options. The uncertainty animation was considered as a useful check by "many radiologists" (a statement that is a bit vague). Similar to Hahn et al. [HPSP01]

and Oeltze et al. [OP05], vessel visualizations were evaluated. However, the tasks are strongly different. While the former publications address *surgeons* operating in the surrounding of *healthy* vessels, Lundstroem et al. address *radiologists* diagnosing *vascular pathologies*. The chosen tasks adequately reflect these differences.

Virtual Endoscopy of the Nasal Region. Krüger et al. [KKPS08] performed the evaluation of SINUSENDOSCOPY a tool for virtual endoscopy in several stages using both formal and summative evaluation. The tool should support preoperative planning, e.g. an identification of patient-specific risks. The focus was on a high degree of realism, e.g. by using textures that simulate wetness. The formal evaluation revealed that the complexity of the user interface to adjust rendering parameters was too complex. In several stages, the number of parameters was strongly reduced and better default values for the remaining parameters were identified. In the formal evaluation, surgeons used the tools for planning surgery and resident surgeons used it as part of a training course.

The summative evaluation was carried out by letting three ENT (ear, nose, throat) surgeons plan 102 surgeries with the tool. They filled questionnaires and commented on the realism (similarity to intraoperative views) and on anatomical structures where the special 3D visualization provided a strong benefit because these structures are highly variable. Some further ideas, e.g. for intraoperative use and more in-depth planning were raised by the users. On the downside details are missing in the description and others are buried in the text instead of being reported in a structured manner.

Exploration of 4D MRI Blood Flow using Stylistic Visualization. vanPelt et al. [vPBB*10] introduced illustrative visualization techniques to convey the essential information from unsteady cardiac blood flow data. Among others, they investigated different seeding strategies to display a representative sample of the flow. Another essential aspect was the display of anatomical context, e.g. the large vascular structures. Illustrative styles and illumination effects are among the parameters that could be adjusted.

To evaluate their method they asked four physicians to fill a questionnaire. Since the pre-clinical cardiovascular blood flow research community was small at that time, four users are indeed appropriate to gather feedback. The evaluation lead to a number of insights. Some parameter adjustments are preferred by all physicians, e.g. the rendering of contours. Some parameter adjustments seem to be dependent on the dataset. A strong aspect of their system is the careful analysis of interaction facilities, e.g. to probe the flow in a standardized way. The discussion clearly shows that standardized measurement planes to obtain quantitative values support a research workflow well. On the other hand, there are also advanced features that seem justified but were not appreciated by any physician. It is important to report such observations as well. Finally, flow speed was assessed as the most important flow feature and therefore its depiction was analyzed. Perhaps surprisingly, the rainbow color scale "was generally valued best to inspect the blood flow speed". It needs to be tested whether they might draw wrong conclusions with their favorite color scale.

Medical Visualization Table. One of the best examples for an evaluation of a medical visualization system was described by Lundstroem et al. [LRF*11] where the medical visualization table with advanced volume rendering and multi-touch-based interaction techniques was introduced (see Fig. 1). The user study is carefully described on three pages. The participants of the study are indeed representative for the target user group: five orthopedic surgeons. They should solve realistic tasks, namely to diagnose a fracture, decide about the treatment and in case of a surgical treatment they should describe their operative strategy in detail. Different methods were combined to collect as much useful feedback as possible. The surgeons were interviewed, a questionnaire was used and "think-aloud" protocols could be analyzed to reveal what participants wanted to achieve and how the system supported them. The study revealed relations between details of the visualization hard- and software and the satisfaction of the participants. Some insights were revealed, e.g. that the display of the patient in their natural size makes it easy to select an appropriate implant. The touch-based interaction was considered as a type of motor learning that supports transfer to surgery. Furthermore, the chances and issues of an intraoperative use could be discussed in detail. Finally, the way the results are presented, is inspiring. As an example, they presented a list of lessons learned.

Figure 1: *The Medical Visualization Table was carefully evaluated with 5 orthopedic surgeons discussing clinically relevant cases (From: [LRF*11]).*

Despite the positive aspects of this evaluation, there is also room for improvement. A "lessons learned" discussion would benefit from a correlation to previous work to better characterize which lessons are really new and which lessons confirm previous research. A second limitation is the short duration of the study. Even as a reader you can grasp how the participants were overwhelmed by the attractive hard- and software design. It would not come as a surprise when after a few weeks of regular use, the benefit would be considered lower. Such changes in the perceived usefulness and attraction can only be identified with long-term evaluations.

BiopsyPlanner. The BiopsyPlanner [HMP*12] supports the process of planning a safe trajectory for acquiring a biopsy in case of brain tumors. Biopsy needles with different diameter are considered. Safety relates to sufficient margins to structures at risk, in particular vascular structures. A number of individual and linked visualizations were developed to be used in coordinated views. For example, a needle pathway distance graph conveys the distance to the closest vessel at every point of the planned trajectory and an enhanced slice

view displays the path, the safety margin in the selected slice and the risk structures. Orthogonal slice views and 3D views enable an in-depth understanding of the entry point, the (linear) path and the exit point.

The evaluation involved five physicians, two of which actually performed biopsies for a long time, whereas the other three have limited experience with biopsies. The physicians should perform a number of predefined tasks, i.e. they used the BiopsyPlanner themselves. Standardized usability questionaires were filled and semi-structured interviews were carried out individually with each expert. Again, a think-aloud protocol was used. Not surprisingly, most in-depth comments were gathered by the two experts that are experienced with these biopsies. This highlights that not only physicians are needed for such evaluations, but that the narrower target user group is required to get substantial feedback. At the same time, this typically restricts the number of participants to a very few. The evaluation are in line with several hypotheses stated in advance. To confirm the hypotheses, of course, a study with more participants is needed. An essential hypothesis is that the planning time was indeed reduced. Advanced computer support may often improve the quality of preoperative decisions often at the expense of additional time which reduces practical feasibility. Moreover, the users stated that they have much more *confidence* to the selected pathway since they can explore such pathways in a more systematic and comprehensive measure. Thus, trust is probably an often useful criterion for diagnostic and treatment planning applications.

PelVis. The PelVis system [SLK*17] aims at surgery planning in case of prostate, rectal and cervic cancer. The complex target anatomy in the pelvic area, the large variability of the relevant structures, their complex spatial relations and the fact that some essential structures are barely visible in image data, are special features for this application. Instead of access or implant planning, the major task here is to define the right resection area to remove the tumor entirely without hurting any risk structure, in particular nerves in the pelvic region. The target structures are segmented from MRI data and combined with information from an atlas to which the patient data is registered. To support the planning process, 3D visualizations with distance encoding, contour renderings as well as linked visualizations were generated.

Five physicians and ten non-domain experts took part in the evaluation. We focus on the reported feedback of the five physicians who are carefully characterized giving a rich picture of the participants. The users could interact with the system themselves. The interview is semi-structured and thus allows the participants to freely describe their impression of the system. A reusable aspect of their evaluation is the general categorization of their questions that relate to the context structure visualization, target structure visualization, risk structure visualization, MRI visualization, and the interaction. A statement that probably holds for most surgical training applications uttered by an expert was “interesting or difficult cases with several pathologies” are particularly relevant. Thus, the case selection is (at least) equally important than details of an advanced visualization technique. Another user reported that the explicit visualization of further anatomical structures is needed. The feedback is reported for each user individually. Thus, the reader may relate the statements to details of the experience of the users. Also, the level of agreement with a number of statements related to the individual visualization techniques introduced here was reported individually for each user instead of a coarse summary statistics.

Summary. Most evaluations relate to the visual encoding and the choice of specific algorithms. These are just two of the four levels of visualization design (recall Munzner [Mun09]). Thus, whether the actual problems in a domain are solved (level 1) or the right abstractions and tasks are used (level 2) is often neglected. A few aspects occured in several evaluations:

Medical visualizations are often assessed w.r.t. their degree of visual realism, i.e. the similarity to intraoperative views. More specifically, the appropriateness for certain diagnostic or therapeutic tasks can be assessed (recall, e.g., [LRF*11]). Do advanced visualizations provide useful information ore even information that directly influences treatment decisions? Evaluations should also consider interactive aspects, not only the quality of rendered images. The evaluation of bloodflow-related visualization techniques from van Pelt et al. [vPBB*10] may serve as a good example for this.

Visualization techniques may have a number of inherent parameters, often too many to be effectively explored. Therefore, evaluations may aim at summarizing parameters, finding appropriate names (in the language of physicians) and default values eventually adapted to the particular data to further support the use of an innovative visualization technique. In the same vein, feedback of several physicians enables an assessment of individual differences which is important to decide about the necessary flexibility.

5. Contributions to a Research Agenda

There are several strategies for a fruitful further research in medical visualization. We do not want to discuss medical visualization subfields that are more or less promising, e.g., based on developments in image acquisition as well as machine learning. Instead, we want to discuss some issues toward a more *relevant* medical visualization research inspired by our analysis of the evaluation practice.

The lack of convincing empirical evaluations probably has a number of reasons. Medical visualization researchers, in contrast to many information visualization researchers, often lack an in-depth understanding of empirical research practices. Eye tracking technology, although affordable, is often not available or the experience to use it efficiently is missing. Thus, there is an educational issue to be solved, e.g., with tutorials. Moreover, the attitude towards evaluation often considers this process a tedious, annoying, but inevitable last minute activity. While we feel that a lack in evaluation is acceptable when presenting new ideas to other researchers, they are of high relevance to practitioners and other readers.

Visualization conferences meanwhile explicitly call for *evaluation papers*. This opportunity should be used by medical visualization researchers much more often. Such papers make it possible to describe comprehensive evaluations with different methods carried out at different stages using a larger set of tasks and datasets to assess the strengths and limitations of a method in more depth than a primarily technical paper could do. Ideally, these evaluation papers arise in another research group than the original technical innovation, since a certain independence is clearly essential to yield

Table 1: *Major empirical evaluations of medical visualization systems.*

Publication	Participants	Eval. methods	Criteria
[OP05]	11	Questionaire	Comprehensibility, similarity to intra-op. views
[BHWB07]	1	Think-aloud	Use of the system for preoperative planning, e.g., resection planning
[KKPS08]	3	Questionaire	Usefulness for surgery planning and patient education, visual realism
[LLPY07]	12	Measurement	Diagnosis and localization of a stenosis
[vPBB*10]	4	Questionaire	Parameter adjustments, color scales
[LRF*11]	5	Think-aloud, interviews, questionaire	Usefulness and usability, strong discussion of interaction
[GNBP11]	2	Interviews	Lens design, e.g., color scales
[HMP*12]	5	Questionaire and semi-structured interview	Planning time, quality of preoperative decisions, confidence in decisions
[SLK*17]	5	Semi-structured interviews	Appropriateness of vis. techniques and case selection

unbiased and trustworthy results. This way, it also becomes possible for the original innovator to publish novel ideas without a request for evaluation hindering the publication of these ideas. We believe that this is of uttermost importance for the visualization community, as it gives other researchers fast access to novel ideas. We see such an idea-first-evaluation-second-strategy also well in line with the concept of post-publication reviews, which become more and more important in many research disciplines embracing an open publication culture. If instead an evaluation dogma would arise in our research community, we fear that the publication of novel and fresh ideas is in danger. As an example, we would like to again discuss the CPR paper published by Kanitsar et al. [KFW*02], which can be considered as a successful example of medical visualization research. Although the basic CPR approach was devoloped earlier and described in a radiology paper [AMRB98], the paper by Kanitsar et al. has spawned several followup publications, and is implemented in medical workstations. While the paper describes an apparently helpful solution to a driving problem, i.e., vessel visualization, the paper does not contain any feedback from the target user group. Thus, the applicability to the problem at hand is only judged by the authors. Nevertheless, the idea has spawned a vast amount of new research, which led to algorithms which today are of high practical relevance, as some form of CPR is realized in most medical workstations. We see this as a valid example from the medical visualization literature, that evaluation requests should not be used as gatekeepers, and thus avoid fresh and novel ideas to be disseminated to the research community early on.

Task taxonomy. One essential contribution, which would help future evaluations, would be a discussion of general medical visualization tasks and their relation to each other. Task taxonomies, such as those of Shneiderman [Shn03], are useful to structure the design and evaluation of future medical visualization techniques. Such a taxonomy would also help to better compare existing approaches. A starting point for establishing such a taxonomy could be to extract common tasks in tumor surgery planning, needle placement, or diagnosis of typical vascular diseases, such as stenoses or plaques. Tumor surgery planning, no matter which organ is affected, often needs a crucial understanding of (possible) infiltrations, the spatial surrounding, and a safety margin. A discussion of typical tasks is also needed for other areas, such as diagnostic virtual endoscopy and medical flow-related diagnosis and treatment. Such a task taxonomy could also lead to reusable questionnaires.

User-centered design. The lack of realistic evaluations is part of a more general problem—developments are largely technically motivated and do not consider actual work practices and constraints. Clinically useful techniques and systems "need to meet extreme demands on simplicity and efficiency" [LLPY07] — a statement that is often overlooked. A more *user-centered* attitude, that is to a large extent a *physician-centered* attitude, is necessary to derive valid requirements that truly reflect user needs instead of introducing tasks visualization researchers want to deal with that are only loosely connected to clinical problems. As an example for the lack of such a user-centered approach, it is notable how often uncertainty related to medical image data is analyzed and visualized (see for instance Ristovski et al. [RPHL14] for a survey). Uncertainty visualization is motivated by the attempt to better support decisions and to avoid wrong conclusions. This is a honorable goal but there are rarely attempts to study how uncertainty visualization actually changes or improves clinical decisions (under the time constraints physicians have in practice). Amar and Stasko (recall [AS04]) discuss the relation between uncertainty and analytical processes. Uncertainty visualizations are often complex, cluttered and somehow fuzzy, whereas decisions are selections from an often small discrete set of options. Thus, physicians have to "de-fuzzify" complex visualizations. Truly, user-centered uncertainty visualizations would be a strong contribution to medical visualization and eventually may be embedded in commercial systems. From this point of view Lundstroem et al. [LLPY07] represent a strong example where clinically relevant tasks were selected and 12 physicians explained how they use the developed techniques.

Collaboration. An essential aspect of professional work in all research areas is the collaboration between colleagues and between experts of different domains. In HCI, many developments aim at supporting collaboration, e.g., recently collaborative Virtual Reality gained a lot of attention. In medicine, physicians of various disci-

plines jointly decide about the treatment of cancer patients in tumor boards, radiologists heavily rely on support by radiology technicians and also surgery is a team effort involving anaesthesiologists, nurses and surgeons. As a consequence, user need assessments, visualization design and later evaluations should consider such professional collaboration aspects.

Touch-based interaction. We discussed the Medical Visualization Table as a successful system with an in-depth evaluation (recall [LRF*11]). A potentially overlooked aspect of this system is its convenient touch-based interaction. Despite problems, such as occlusions of the target by touch input and a lack of precision, the increased directness of touch input is very promising. Multitouch interfaces naturally support bimanual interaction, a type of interaction that is useful for tasks like object placement or scaling. Carefully designed touch-based interaction outperforms the traditional mouse and keyboard input for many tasks and is also potentially useful for 3D interaction tasks (3D selection, translation and rotation), as they are important for medical visualization systems [BF07, HCC07, HtCC09]. Touch-based interaction with table-like output devices may provide improved user interaction, in particular collaborative user interaction and contribute to a broader use of advanced visualization techniques. Research tasks may include designing optimal virtual anatomy or surgical planning tools operated by multitouch input.

Again we would like to stress that we do not argue that each medical visualization paper needs a comprehensive evaluation. There are convincing examples where the need for new visualization techniques is carefully analyzed and where the new technique is illustrated with a representative selection of pathological cases and thus the value of the introduced technique is self-evident (recall [LWP*06]). Also, more recently, Kretschmer et al. [KST*14]) came up with "Anatomy-Driven Reformation" and used clinically relevant examples to demonstrate their work without any user feedback. We as visualization researchers and reviewers should be able to identify such approaches and value their novelty. When the visualization design is convincingly justified, e.g., w.r.t. known perceptual rules or established models, an evaluation is not necessary.

6. Conclusion

Medical visualization has a number of success stories. Early work on CPR vessel visualization (and some later refinements, e.g., for multi-path vessels) and virtual endoscopy have found their way in radiological workstations. Advanced 3D vessel visualization techniques were incorporated in liver surgery planning and extensively used in this area. Some systems and techniques were at least used by *one* clinical partner who co-developed the system. Though, the large majority of techniques introduced later did not experience any use outside the labs they were developed in. There are potentially different reasons for this situation: visualization techniques were developed without a deep understanding of the problems they should help to solve, the potential clinical impact had a lower priority compared to embellishments that make a technique attractive for the audience of a visualization conference. However, also the lack of realistic evaluations might contribute to the growing gap between medical visualization research and actual use of visualization techniques in medical research and clinical practice.

The current situation in medical visualization is comparable to the situation in information visualization before the BELIV workshop series started in 2006. The medical visualization research community can learn a lot from the research presented at these workshops and at recent VAST and InfoVis conferences. Very likely, an increased focus on in-depth evaluations leads to new ideas for further research, ideas that are closer to real clinical needs than many ideas realized in the past. In this sense, evaluation may even be considered as a creativity technique.

With this paper, we argue for a stronger focus on evaluation in medical visualization taking the whole spectrum of qualitative and quantitative methods applied at different stages into account. In particular, we emphasize the need for evaluations in realistic settings in contrast to lab-based experiments using overly simple tasks and computer science students to "replace" the actual target user group of medical visualization systems. When developing a single visual representation mainly focusing on perceptual aspects, this is often sufficient. However, when developing a medical visualization system serving a targeted medical workflow, most likely an expert user is needed to asses the situation – either by contributing to the design or by taking part in the evaluation.

When an expert evaluation is necessary, the specific evaluation strategy should be discussed with physicians, nurses, radiology technicians, medical researchers or students of medicine. Realistic tasks and representative cases, including cases where the diagnosis and treatment planning is difficult, should be prepared for empirical evaluations. Eye-tracking and think-aloud may enhance the understanding of how physicians use visualization technology. Long-term case studies should be taken into account to study the usage of interactive visualization techniques more deeply.

Our critical analysis is limited by the papers that were selected for our discussion, i.e., primarily papers published at IEEE VIS and EuroVis. Although we tried to incorporate all major subfields of medical visualization, there is a small risk that we have overlooked an essential aspect of the evaluation practice.

Acknowledgements. We thank Helwig Hauser, University of Bergen as well as Christian Hansen and Benjamin Köhler, Univ. of Magdeburg for commenting on the manuscript and acknowledge the Dagstuhl Seminar 18041 "Foundations of Visualization" where evaluation issues were carefully discussed—stimulating this manuscript.

References

[AA16] AZER S. A., AZER S.: 3d anatomy models and impact on learning: a review of the quality of the literature. *Health professions education 2*, 2 (2016), 80–98. 4

[AMRB98] ACHENBACH S., MOSHAGE W., ROPERS D., BACHMANN K.: Curved multiplanar reconstructions for the evaluation of contrast-enhanced electron beam ct of the coronary arteries. *American journal of roentgenology 170*, 4 (1998), 895–99. 9

[And08] ANDREWS K.: Evaluation comes in many guises. In *Proc. of AVI Workshop on BEyond time and errors (BELIV)* (2008), pp. 7–8. 2

[AS04] AMAR R., STASKO J.: A knowledge task-based framework for design and evaluation of information visualizations. In *IEEE Symposium on Information Visualization* (2004), pp. 143–50. 1, 4, 9

[BF07] BENKO H., FEINER S. K.: Balloon selection: A multi-finger technique for accurate low-fatigue 3d selection. In *Proc. of IEEE Symposium on 3D User Interfaces* (2007), pp. 22–29. 10

[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 Trans. Vis. Comput. Graph. 17*, 12 (2011), 2479–88. 4

[BHW*07] BURNS M., HAIDACHER M., WEIN W., VIOLA I., GRÖLLER E.: Feature Emphasis and Contextual Cutaways for Multimodal Medical Visualization. In *Proc. EuroVis* (2007), pp. 275–82. 4

[BHWB07] BEYER J., HADWIGER M., WOLFSBERGER S., BÜHLER K.: High-quality multimodal volume rendering for preoperative planning of neurosurgical interventions. *IEEE Trans. Vis. Comput. Graph. 13*, 6 (2007), 1696–1703. 6, 9

[BKR*14] BLASCHECK T., KURZHALS K., RASCHKE M., BURCH M., WEISKOPF D., ERTL T.: State-of-the-art of visualization for eye tracking data. In *Proc. of EuroVis* (2014). 3

[BWKG01] BARTROLÍ A. V., WEGENKITTL R., KÖNIG A., GRÖLLER E.: Nonlinear Virtual Colon Unfolding. In *Proc. IEEE Visualization* (2001), pp. 411–20. 4

[Car08] CARPENDALE S.: Evaluating information visualizations. In *Information visualization*. Springer, 2008, pp. 19–45. 1, 2

[GNBP11] GASTEIGER R., NEUGEBAUER M., BEUING O., PREIM B.: The FLOWLENS: A focus-and-context visualization approach for exploration of blood flow in cerebral aneurysms. *IEEE Trans. Vis. Comput. Graph. 17*, 12 (2011), 2183–92. 5, 9

[HCC07] HANCOCK M. S., CARPENDALE S., COCKBURN A.: Shallow-depth 3d interaction: design and evaluation of one-, two- and three-touch techniques. In *Proc. of the ACM SIGCHI conference on Human Factors in Computing Systems* (2007), pp. 1147–56. 10

[HMK*97] HONG L., MURAKI S., KAUFMAN A. E., BARTZ D., HE T.: Virtual voyage: interactive navigation in the human colon. In *Proc. of ACM SIGGRAPH* (1997), pp. 27–34. 4

[HMP*12] HERGHELEGIU P. C., MANTA V., PERIN R., BRUCKNER S., GRÖLLER E.: Biopsy planner - visual analysis for needle pathway planning in deep seated brain tumor biopsy. *Comput. Graph. Forum 31*, 3 (2012), 1085–94. 6, 7, 9

[HPSP01] HAHN H. K., PREIM B., SELLE D., PEITGEN H.: Visualization and interaction techniques for the exploration of vascular structures. In *Proc. IEEE Visualization* (2001), pp. 395–402. 6

[HQK06] HONG W., QIU F., KAUFMAN A. E.: A pipeline for computer aided polyp detection. *IEEE Trans. Vis. Comput. Graph. 12*, 5 (2006), 861–68. 4

[HS10] HWANG W., SALVENDY G.: Number of people required for usability evaluation: the 10±2 rule. *Communications of the ACM 53*, 5 (2010), 130–33. 2

[HtCC09] HANCOCK M. S., TEN CATE T., CARPENDALE S.: Sticky tools: full 6dof force-based interaction for multi-touch tables. In *Proc. of ACM International Conference on Interactive Tabletops and Surfaces ITS* (2009), pp. 133–40. 10

[HZP*14] HANSEN C., ZIDOWITZ S., PREIM B., STAVROU G., OLDHAFER K. J., HAHN H. K.: Impact of model-based risk analysis for liver surgery planning. *International journal of computer assisted radiology and surgery 9*, 3 (2014), 473–480. 4

[IIC*13] ISENBERG T., ISENBERG P., CHEN J., SEDLMAIR M., MÖLLER T.: A systematic review on the practice of evaluating visualization. *IEEE Trans. Vis. Comput. Graph. 19*, 12 (2013), 2818–27. 1

[JSV*08] JOSHI A., SCHEINOST D., VIVES K. P., SPENCER D. D., STAIB L. H., PAPADEMETRIS X.: Novel interaction techniques for neurosurgical planning and stereotactic navigation. *IEEE Trans. Vis. Comput. Graph. 14*, 6 (2008), 1587–94. 4

[KFW*02] KANITSAR A., FLEISCHMANN D., WEGENKITTL R., FELKEL P., GRÖLLER E.: CPR - curved planar reformation. In *Proc. IEEE Visualization* (2002), pp. 37–44. 4, 9

[KGP*13] KÖHLER B., GASTEIGER R., PREIM U., THEISEL H., GUTBERLET M., PREIM B.: Semi-Automatic Vortex Extraction in 4D PC-MRI Cardiac Blood Flow Data using Line Predicates. *IEEE Trans. Vis. Comput. Graph. 19*, 12 (2013), 2773–82. 5

[KKPS08] KRUEGER A., KUBISCH C., PREIM B., STRAUSS G.: Sinus endoscopy-application of advanced gpu volume rendering for virtual endoscopy. *IEEE Trans. Vis. Comput. Graph. 14*, 6 (2008), 1491–98. 6, 7, 9

[Kru96] KRUPINSKI E. A.: Visual scanning patterns of radiologists searching mammograms. *Academic radiology 3*, 2 (1996), 137–44. 3

[KST*14] KRETSCHMER J., SOZA G., TIETJEN C., SÜHLING M., PREIM B., STAMMINGER M.: ADR - anatomy-driven reformation. *IEEE Trans. Vis. Comput. Graph. 20*, 12 (2014), 2496–2505. 10

[KWFG03] KANITSAR A., WEGENKITTL R., FLEISCHMANN D., GRÖLLER E.: Advanced curved planar reformation: Flattening of vascular structures. In *Proc. IEEE Visualization* (2003). 4

[LBI*12] LAM H., BERTINI E., ISENBERG P., PLAISANT C., CARPENDALE S.: Empirical studies in information visualization: Seve scenarios. *IEEE Trans. Vis. Comput. Graph. 18*, 9 (2012), 1520–36. 1, 2

[Lew12] LEWIS J. R.: *Usability Testing*. Wiley-Blackwell, 2012, ch. 46, pp. 1267–1312. 2

[LFH17] LAZAR J., FENG J., HOCHHEISER H.: *Research Methods in Human-Computer Interaction*. Morgan Kaufmann, 2017. 1

[LGF*00] LAMADÉ W., GLOMBITZA G., FISCHER L., CHIU P., CÁRDENAS SR C. E., THORN M., MEINZER H.-P., GRENACHER L., BAUER H., LEHNERT T.: The impact of 3-dimensional reconstructions on operation planning in liver surgery. *Archives of surgery 135*, 11 (2000), 1256–1261. 4

[LLPY07] LUNDSTRÖM C., LJUNG P., PERSSON A., YNNERMAN A.: Uncertainty Visualization in Medical Volume Rendering Using Probabilistic Animation. *IEEE Trans. Vis. Comput. Graph. 13*, 6 (2007), 1648–55. 6, 9

[LRF*11] LUNDSTRÖM C., RYDELL T., FORSELL C., PERSSON A., YNNERMAN A.: Multi-Touch Table System for Medical Visualization: Application to Orthopedic Surgery Planning. *IEEE Trans. Vis. Comput. Graph. 17*, 12 (2011), 1775–84. 1, 6, 7, 8, 9, 10

[LWP*06] LJUNG P., WINSKOG C., PERSSON A., LUNDSTROM C., YNNERMAN A.: Full body virtual autopsies using a state-of-the-art volume rendering pipeline. *IEEE Trans. Vis. Comput. Graph. 12*, 5 (2006), 869–76. 4, 10

[MMV*13] MISTELBAUER G., MORAR A., VARCHOLA A., SCHERNTHANER R., BACLIJA I., KÖCHL A., KANITSAR A., BRUCKNER S., GRÖLLER E.: Vessel visualization using curvicircular feature aggregation. *Comput. Graph. Forum 32*, 3 (2013), 231–240. 5

[MSL*18] MEUSCHKE M., SMIT N., LICHTENBERG N., PREIM B., LAWONN K.: Automatic Generation of Web-Based User Studies to Evaluate Depth Perception in Vascular Surface Visualizations . In *Proc. VCBM* (2018). 3

[Mun09] MUNZNER T.: A nested model for visualization design and validation. *IEEE Trans. Vis. Comput. Graph. 15*, 6 (2009). 1, 2, 8

[NO99] NARDI B. A., O'DAY V.: *Information ecologies: Using technology with heart*. Mit Press, 1999. 3

[Nor06] NORTH C.: Toward Measuring Visualization Insight. *IEEE CG&A 26*, 3 (2006), 6–9. 1

[OP05] OELTZE S., PREIM B.: Visualization of vasculature with convolution surfaces: method, validation and evaluation. *IEEE Transactions on Medical Imaging 24*, 4 (2005), 540–48. 6, 7, 9

[PBC*16] PREIM B., BAER A., CUNNINGHAM D., ISENBERG T., ROPINSKI T.: A survey of perceptually motivated 3d visualization of medical image data. *Comput. Graph. Forum 35*, 3 (2016), 501–25. 3, 4

[Pla04] PLAISANT C.: The challenge of information visualization evaluation. In *Proc. of Advanced visual interfaces* (2004), pp. 109–16. 1

[PS18] PREIM B., SAALFELD P.: A survey of virtual human anatomy education systems. *Computers & Graphics 71* (2018), 132–153. 4

[PWN*15] PANKAU T., WICHMANN G., NEUMUTH T., PREIM B., DIETZ A., STUMPP P., BOEHM A.: 3d model-based documentation with the tumor therapy manager (ttm) improves tnm staging of head and neck tumor patients. *International journal of computer assisted radiology and surgery 10*, 10 (2015), 1617–1624. 4

[RHD*06] RITTER F., HANSEN C., DICKEN V., KONRAD-VERSE O., PREIM B., PEITGEN H.: Real-time illustration of vascular structures. *IEEE Trans. Vis. Comput. Graph. 12*, 5 (2006), 877–84. 2, 4

[RHR*09] ROPINSKI T., HERMANN S., REICH R., SCHÄFERS M., HINRICHS K. H.: Multimodal vessel visualization of mouse aorta PET/CT scans. *IEEE Trans. Vis. Comput. Graph. 15*, 6 (2009), 1515–22. 4, 5

[RPHL14] RISTOVSKI G., PREUSSER T., HAHN H. K., LINSEN L.: Uncertainty in medical visualization: Towards a taxonomy. *Computers & Graphics 39* (2014), 60–73. 9

[RRRP08] RIEDER C., RITTER F., RASPE M., PEITGEN H.: Interactive Visualization of Multimodal Volume Data for Neurosurgical Tumor Treatment. *Comput. Graph. Forum 27*, 3 (2008), 1055–62. 4, 5

[RSH06] ROPINSKI T., STEINICKE F., HINRICHS K. H.: Visually supporting depth perception in angiography imaging. In *Proc. Smart Graphics* (2006), pp. 93–104. 4

[RvdHD*15] RAIDOU R. G., VAN DER HEIDE U. A., DINH C. V., GHOBADI G., KALLEHAUGE J., BREEUWER M., VILANOVA A.: Visual analytics for the exploration of tumor tissue characterization. *Comput. Graph. Forum 34*, 3 (2015), 11–20. 5

[SHM09] STULL A. T., HEGARTY M., MAYER R. E.: Getting a handle on learning anatomy with interactive three-dimensional graphics. *Journal of Educational Psychology 101*, 4 (2009), 803. 4

[Shn03] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *The Craft of Information Visualization*. Elsevier, 2003, pp. 364–71. 9

[Shn10] SHNEIDERMAN B.: *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India, 2010. 1

[SK10] SCHULTZ T., KINDLMANN G. L.: Superquadric glyphs for symmetric second-order tensors. *IEEE Trans. Vis. Comput. Graph. 16*, 6 (2010), 1595–1604. 2

[SLB*18] SAALFELD P., LUZ M., BERG P., PREIM B., SAALFELD S.: Guidelines for quantitative evaluation of medical visualizations on the example of 3d aneurysm surface comparisons. *Comput. Graph. Forum 37* (2018). 3

[SLK*17] SMIT N. N., LAWONN K., KRAIMA A., DERUITER M., SOKOOTI H., BRUCKNER S., EISEMANN E., VILANOVA A.: Pelvis: Atlas-based surgical planning for oncological pelvic surgery. *IEEE Trans. Vis. Comput. Graph. 23*, 1 (2017), 741–50. 6, 8, 9

[SP06] SHNEIDERMAN B., PLAISANT C.: Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proc. of the Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization* (2006). 3

[SzBBKN14] SCHULTE ZU BERGE C., BAUST M., KAPOOR A., NAVAB N.: Predicate-based focus-and-context visualization for 3d ultrasound. *IEEE Trans. Vis. Comput. Graph. 20*, 12 (2014), 2379–87. 5

[TM04] TORY M., MÖLLER T.: Human factors in visualization research. *IEEE Trans. Vis. Comput. Graph. 10*, 1 (2004), 72–84. 2

[VBvdP04] VILANOVA A., BERENSCHOT G., VAN DE PUL C.: DTI visualization with streamsurfaces and evenly-spaced volume seeding. In *Data Visualization (Proc. Eurographics/IEEE Symposium on Visualization)* (2004), pp. 173–82. 4

[vPBB*10] VAN PELT R., BESCÓS J. O., BREEUWER M., CLOUGH R. E., GRÖLLER E., TER HAAR ROMENY B. M., VILANOVA A.: Exploration of 4d MRI blood flow using stylistic visualization. *IEEE Trans. Vis. Comput. Graph. 16*, 6 (2010), 1339–47. 6, 7, 8, 9

[WS96] WHITTAKER S., SIDNER C.: Email Overload: Exploring Personal Information Management of Email. In *Proc. of ACM SIGCHI Conference on Human Factors in Computing Systems* (1996), pp. 276–83. 3