Research Design Considerations for Empirical User Studies in MICCAI

Sue Min Cho[1⋆], Catalina Gomez[1⋆], Katharina Breininger[2], Francis Creighton[3], Xiaoqing Guo[4], Dean Ho[5], Masaru Ishii[3], Pierre Jannin[6], Marta Kersten[7], Seong Tae Kim[8], Nassir Navab[9], Cheng Ouyang[10], Shandong Wu[11], Paul Yi[12], Maria A. Zuluaga[13], and Mathias Unberath[1]

[1] Johns Hopkins University, USA
[2] Universität Würzburg, Germany
[3] Johns Hopkins Medicine, USA
[4] Hong Kong Baptist University, Hong Kong, China
[5] National University of Singapore, Singapore
[6] Inserm Université de Rennes, France
[7] Concordia University, Canada
[8] Kyung Hee University, South Korea
[9] Technical University of Munich, Germany
[10] University of Oxford, UK
[11] University of Pittsburgh, USA
[12] St. Jude Children's Research Hospital, USA
[13] EURECOM, France
{scho72, unberath}@jhu.edu

Abstract. As technologies for applications in medical imaging and computer-assisted interventions mature, technical feasibility alone is insufficient for translation to bedside. Highlighted by communities that emphasize human-centered technology design and recent efforts at MICCAI, there is a need to evaluate technological solutions with human involvement. Among the various forms of human-centered evaluation, we focus on empirical, controlled user studies of human-technology interaction. These evaluations involve actual applications and real end-users, providing actionable evidence about human impact and intended use. However, the lack of standardized studies and unique challenge in user studies complicate the measurement of progress and reproducibility in research. This paper aims to support the MICCAI community in conducting evaluations with human subjects, by drawing on insights from human-centered principles and quantitative research in human-computer interaction. We discuss key considerations that include the formulation, planning, and execution of valid and meaningful experiments. The rigor and use of systematic methods in human-subjects evaluations will ultimately provide researchers with more reliable evidence to assert the potential achievements of their technological solutions. As considerations for human-subject experiments differ from those for technical development, the evaluation of these works demands criteria within the MICCAI

⋆ Equal contribution

community that account for the substantial effort and rigor required to conduct human-subject evaluations. Findings and lessons from human-machine interaction studies will guide future research aimed at deploying technologies and assessing clinical effectiveness.

Keywords: Human-subjects research · human-in-the-loop · human-machine interaction · controlled user study

1 Introduction

The integration of computational systems in medicine introduces technology designed to assist with clinical tasks, from diagnostic image interpretation to surgical guidance and therapeutic interventions. Understanding the most effective form of technological assistance for these tasks is crucial and requires evaluations contextualized within specific use cases or application-grounded scenarios [6]. Most research in computing systems focuses on technical metrics to demonstrate that algorithms perform accurately and reliably in solving clinical tasks [21]. However, these evaluations neglect measures of human-technology interaction. This is problematic when integrating medical image computing (MIC) and computer-assisted intervention (CAI) technologies into clinical settings, where human users will interact with these systems and use them to assist critical decisions or actions within complex workflows [23].

Human-centered evaluation methods include intentional assessments of human performance and experience [17]. In particular, empirical evaluations provide a testbed for studying the interactions between humans and technologies as they complete relevant tasks, helping to understand the factors that shape users' behavior, expectations, and outcomes of the collaboration. By conducting structured tests with representative users, tasks, and context, this approach generates empirical evidence that supports and justifies algorithmic designs. The rigor and use of systematic methods in human-subjects evaluations of MIC and CAI solutions will ultimately provide the evidence for researchers to claim potential achievements of these technologies, reduce unexpected failures and negative impact of AI, and move toward real-world evaluation [2].

The timing of controlled user studies in MIC and CAI systems development remains debated. Some prioritize algorithm design and performance optimization before human evaluation [20, 3, 19], while others incorporate iterative user input earlier, often through simulations or prototypes [10, 13, 4]. Despite these differing approaches, both ultimately require empirical human-subject studies to assess the system's impact on user behavior and ensure safe progression toward real-world implementation [18].

The pursuit of human-subjects evaluations across MIC and CAI invites us to reexamine and evolve beyond the conventional research and development paradigms traditionally embraced by the MICCAI community. As we venture into this relatively uncharted domain, we need to cultivate a shared understanding and support its added value to current benchmarks. In Section 2, we set the stage for a deeper exploration into considerations for controlled user studies.

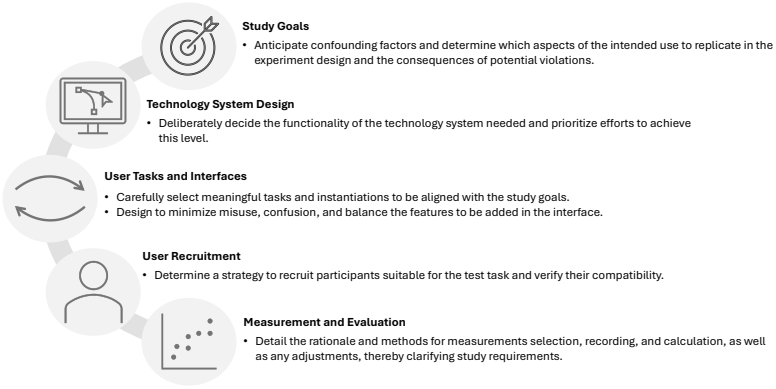

Fig. 1. Considerations in Controlled User Studies in MIC and CAI Applications.

Following this, Section 3 calls for action to the MICCAI community to foster a culture of openness and to adopt conscientious and best practices in support of these approaches within human-centered evaluation. Through these sections, the paper aims to spark a conversation within the MICCAI community about embracing this critical pursuit and to outline a collective and systematic approach towards the development of more innovative, usable, and practical MIC and CAI solutions that move from bench to bedside.

2 Empirical User Study Design Considerations across MIC and CAI

We present a set of practical considerations, summarized in Figure 1 and illustrated in a case study (Sec. 2.6), that inform the formulation, planning, and execution of valid and meaningful controlled user studies and promote good practices.

2.1 Identifying Study Goals

User studies with MIC and CAI technologies resemble psychology and human-computer interaction (HCI) testing. They require human participants whose experiences are studied either in controlled laboratory conditions or in real-world contexts. Evidence collected through these studies can inform researchers about potential achievements of the technological solutions they are envisioning, rather than relying on assumptions. Research in explainable AI (XAI) has particularly emphasized this perspective by challenging a purely computational approach to developing explainable mechanisms and advocating for the integration of human-centered practices [4]. As with other types of empirical evaluations, researchers establish if a change in an experimental condition causes a change in an outcome

measurement, resulting in a hypothesis that can be tested through controlled experiments or refined through exploratory research [15]. The validity of the results can be assessed internally and externally. Internal validity refers to the aspects of a study's design that ensure that the causal relationships are valid, specifically by reducing confounding factors. External validity considers the aspects that capture the broader applicability of the results to additional groups of people, contexts, etc. [22]. The challenge of balancing experimental control with ecological realism is central to designing studies that are both rigorous and clinically relevant [14, 5]. *To account for both, anticipate confounding factors and determine which aspects of the intended use to replicate in the experiment design and the consequences of potential violations.*

Formulating feasible experimental conditions where participants can engage with a technology is a pivotal aspect of study design that requires planning interventions. Successful implementation and validation of these manipulations are key to ensuring the study's internal validity. This process involves not only devising interventions that can be realistically applied within the study's constraints (e.g., equipment, ethical approval, participant, and data availability) but also establishing reliable methods to evaluate their successful execution.

2.2 Technology System Design

In experimental interventions in human-technology interaction studies, researchers must determine both the system design and its execution during the study, selecting the appropriate level of implementation fidelity – from low-fidelity mock-ups to fully functioning systems – based on the research questions being addressed.

Prototype mock-ups serve as a strategic tool for rapidly exploring design opportunities and probing user behaviors in a controlled yet flexible environment before committing to actual implementations [24]. While feasible for minor adjustments in inputs or outputs, mock-ups present limitations when simulating complex computational processes involving medical images.

An intermediate approach involves curating samples of real system behavior, carefully chosen to facilitate the study of specific phenomena and avoid dealing with variability in system outputs [24]. This strategy can still assess risks of errors in diagnostic algorithms [11], segmentation tools, or registration accuracy.

Unlike mock-ups or curated samples, a fully functioning system can process inputs, adapt to user actions, and provide appropriate responses, offering more dynamic interactions, real-time feedback, and authentic experience. However, modifying a fully implemented system based on user feedback can be slow and resource-intensive, hindering rapid iteration. When rapid iteration is a priority, researchers can use conceptual prototypes that incorporate real system outputs, provided that the functionalities under study are grounded in established technical concepts.

When physical interaction with the environment is involved, design considerations must go beyond screen-based interfaces to include physical systems. Surgical navigation systems, robotic platforms, and image-guided tools require

careful integration of hardware and software. Simulation environments and surgical phantoms allow researchers to study human interaction without risk, but simulation fidelity affects how well findings translate to clinical use [8]. Different study types – benchtop, phantom, cadaver, and in-vivo – vary in realism and approval requirements, making iterative design more challenging due to potential hardware modifications. *In general, deliberately decide the functionality needed for achieving study goals, then prioritize efforts accordingly.*

2.3 User Tasks and Interfaces

In task-based evaluations, participant activities reflect real-world uses of the tool, system, or product. Tasks may either be well-established ones that are commonly studied and have driven collaborative efforts in computational challenges, or newly designed tasks to showcase functionalities for applications not yet seen in practice. New tasks may emerge as technologies evolve, necessitating development in close collaboration with stakeholders.

Selecting task instances typically involves real data to preserve clinical realism, spanning diverse conditions (e.g., multi-class cases, imbalance, rare events) and avoiding artificial stimuli. Although variety reflects real-world complexity, factors must remain manageable given available resources. Study length influences participant commitment and data quality, and repeated exposure to multiple cases may introduce learning effects that confound the intervention's impact. *Carefully select meaningful tasks and instantiations to be aligned with the study goals.*

Along with tasks, user interface (UI) design shapes participant interaction with the system. UI development should be integral from the beginning of the design process. Early collaboration with experts in HCI, human factors, design, and other relevant areas ensures that the UI effectively supports specific human-technology interaction scenarios, emphasized in key guidelines [12, 1, 16]. Overly complex interfaces can obscure innovative features and complicate assessment, while overly simplistic designs may undermine clinical relevance. *Design to minimize misuse, confusion, and balance the features to be added in the interface.*

Running a pilot study before full data collection is strongly recommended. Pilots help identify usability issues, ambiguities in task instructions, timing problems, and other practical concerns that may not be apparent during design. Feedback from pilot participants allows researchers to refine the study protocol, interface, and procedures before committing to the full study.

2.4 User Recruitment

Evaluations involving human subjects to perform real tasks with the target application demand precise identification and well-thought-out recruitment strategies of participants. Criteria for inclusion must define representativeness and include reliable verification methods. In the medical context, participant samples frequently come from specialized populations with more challenges to access.

Online platforms like Prolific[14] enable recruitment of physicians for surveys. However, although online studies offer broader reach and limited environmental control, participant engagement may threaten external validity, as behavior may not reflect real-world responses. Increasing sample size is not always a remedy: it raises costs and can amplify trivial effects, while fundamental design flaws remain unaddressed. Conducting a priori power analysis is thus strongly recommended to estimate the sample size required to detect a meaningful effect at a chosen significance level, given expected variability [25]. Although baseline human-factor data and effect sizes are scarce in medical domains, this challenge creates an opportunity to adopt more rigorous and validity-enhancing practices.

Variability in participant skills, inherent in both online and in-person recruitment and possibly intentional [7, 10], will naturally occur in experiments with human subjects. Between-subject designs are sensitive to these individual differences and typically require larger samples per condition. To address these individual differences and observe the expected effects with smaller sample sizes – an important consideration when qualified participants may be difficult to recruit – within-subject designs can be used.

Training materials help align participants' skills and knowledge with the study requirements, mitigating task success variability and balancing informed participation with the learning curve. Acknowledging potential end users' lack of training and familiarity with these tools is crucial for interpreting study results, as seeing real benefits may require time for users to become accustomed to the tool, which is a process best captured through long-term observational or field studies, rather than single sessions in a controlled experiment. *In line with study design and implementation, determine a strategy to recruit participants suitable for the test task and verify their compatibility, considering factors such as experience level, technical proficiency, clinical site, and geographic context.*

2.5 Measurement and Evaluation

The research question and goals set different human-desirable constructs that might be of interest and meaningful in determining the success of MIC and CAI technologies. Common constructs include task performance (e.g., accuracy, completion time), cognitive load, trust and reliance, usability, and situational awareness. Established instruments such as the NASA Task Load Index and System Usability Scale can be applied, though many constructs lack validated instruments specific to MIC and CAI contexts.

Defining the construct under investigation and implementing instruments that measure it accurately are the basis to make correct inferences that otherwise would not be fixed by having rigorous study design, advanced analysis methods, or large samples. If the outcomes are not measured using relevant metrics and under certain study designs, no reliable statements or claims can be made about how suitable is the system to achieve certain human engineering goals.

[14] https://www.prolific.com/

Information on how measures were selected and used is essential to assess a study's validity. In psychology, lack of such transparency has been shown to undermine confidence in measures and findings, often prompting reconsideration of claims [9]. Measurement decisions should therefore align with the constructs of interest from early stages, and further work is needed to adapt and validate instruments for MIC and CAI contexts. *Detail the rationale and methods for measurements selection, recording, and calculation, as well as any adjustments, thereby clarifying study requirements.*

2.6 Case Study

To demonstrate practical utility, we examined how a CAI publication on virtual reality simulation for microsurgery by You *et al.* [26] applied these considerations in practice. For study goals (Sec. 2.1), the paper tests the efficiency of two hands-free interaction concepts for controlling visualization adjustments during operative tasks, comparing a simulated setup in virtual reality to a physical session with surgical microscopes as a reference. For technology system design (Sec. 2.2), the system used existing technologies (head-mounted display and responsive virtual screens) to adjust the field of view using head pose and gaze direction data. For user tasks and interfaces (Sec. 2.3), participants completed repeated target selection tasks, simulating operative adjustments requiring visualization changes. For user recruitment (Sec. 2.4), the study targeted neurosurgeons (attendings and residents) as representative end users. For measurement and evaluation (Sec. 2.5), the study recorded completion time and head motion path length as objective performance metrics.

3 Call for Collective, Proactive Action

Recognize the Challenge and Efforts. As illustrated in the previous section, empirical user studies in MIC and CAI involve numerous critical decisions, from defining study goals and obtaining ethical approval to analyzing data. Researchers must determine system fidelity, choose between functional implementations and mock-ups, design and test interfaces, conduct manipulation checks, secure ethics committee approval, recruit participants, set up experimental environments (including specialized equipment or simulations), run pilots, and select appropriate statistical analyses. These iterative steps highlight the substantial complexity and commitment required. Although published studies may appear straightforward, user research demands intensive planning and execution, distinct from purely technical work.

However, these best practices must be balanced against real-world constraints. Pilot studies, though essential, extend timelines and require additional recruitment. Recruiting clinically qualified participants poses significant logistical challenges: coordinating with hospitals, accommodating clinical schedules, and obtaining site-specific approvals can delay studies by months. High-fidelity simulations demand substantial investment in equipment and expertise. Researchers

should prioritize considerations based on their research questions. Early-stage concept validation may reasonably use lower-fidelity prototypes, while claims of clinical readiness require more rigorous ecological validity. Explicitly acknowledging trade-offs strengthens rather than weakens a study's contribution.

Despite these constraints, early empirical investigations, even at a preliminary stage, can yield critical insights into the practicality and effectiveness of these technologies in clinical contexts. These efforts can help identify potential pitfalls of real applications early on, offering opportunities for refinement and adjustment before proceeding to more detailed and committed system development. This is particularly valuable given the safety-critical nature of many MIC and CAI applications, where understanding human factors before deployment can prevent adverse outcomes. We should proactively work to equip our community with the necessary tools, training, and recognition to successfully foster the growth of this area.

Promote Transparency, Openness, and Reproducibility. Transparency in reporting empirical user study methodologies and findings is crucial. This transparency extends beyond detailing the study's design to include justifications for methodological choices and an honest assessment of how well the experimental scenario reflects real-world conditions. For studies involving physical interactions or interventional technologies, this includes reporting on the fidelity of simulation environments, the realism of phantoms or test setups, and any constraints imposed by safety or regulatory requirements. Any deviations from intended clinical use should be clearly stated, not left to interpretation, and acknowledged as limitations. This ensures that the research community and practitioners can accurately evaluate the study's relevance and impact, laying a solid foundation for the systematic growth of human-centered research in MIC and CAI. Furthermore, clear and comprehensive reporting is essential for addressing ethical considerations and institutional review board approval. To support reproducibility, key reporting elements include study goals and hypotheses, participant characteristics and recruitment criteria, task and stimuli design, system fidelity and functionality, interface features, experimental procedure and conditions, measurement instruments and constructs, analysis approach, and limitations.

Open-source code repositories have become standard for technical studies. Empirical user studies could benefit from similar practices to improve reproducibility and knowledge sharing. This includes sharing study protocols, task descriptions, interface designs, questionnaires, and analysis scripts. Pre-registering study designs and hypotheses can enhance the credibility of findings by distinguishing confirmatory analyses from exploratory ones. While sharing raw participant data may be limited due to privacy concerns, providing anonymized or aggregated data, along with documentation of data collection methods, enables researchers to build on existing work, replicate findings, or conduct meta-analyses. Establishing community norms for these sharing practices would help accelerate progress and minimize redundant efforts.

Fostering an accepting and open-minded stance toward human-centered research in MIC and CAI is essential. Given the nascent nature of human-centered

approaches in these fields, it is important to recognize that conventional evaluation metrics and approaches may not always apply. There should be caution against dismissing research based on preconceived notions of valid methodology or impact. Evaluations should be based on evidence, ensuring feedback constructively contributes to refining and advancing the field.

4 Conclusion

This paper has outlined considerations for conducting rigorous controlled user studies in MIC and CAI research, spanning study goals, technology system design, tasks and interfaces, participant recruitment, and measurement. The unique challenges of these domains – specialized user populations, safety-critical applications, and complex equipment – demand careful attention and expertise distinct from purely technical work. Of course, controlled user studies have inherent limitations: findings from laboratory settings may not generalize to real clinical environments, participants may behave differently when observed, and short-term studies cannot capture how adoption unfolds over time. These limitations do not diminish the value of controlled studies, but they highlight the need for complementary approaches such as field studies, longitudinal evaluations, and qualitative inquiry as technologies mature toward deployment. By establishing rigorous practices, the MICCAI community can build a cumulative evidence base that reliably supports claims about human-centered technologies.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., et al.: Guidelines for human-ai interaction. In: Proceedings of the 2019 chi conference on human factors in computing systems. pp. 1–13 (2019)
2. Azad, T.D., Krumholz, H.M., Saria, S.: Principles to guide clinical ai readiness and move from benchmarks to real-world evaluation. Nature Medicine pp. 1–3 (2026)
3. Chanda, T., Hauser, K., Hobelsberger, S., Bucher, T.C., Garcia, C.N., Wies, C., Kittler, H., Tschandl, P., Navarrete-Dechent, C., Podlipnik, S., et al.: Dermatologist-like explainable ai enhances trust and confidence in diagnosing melanoma. Nature Communications **15**(1), 524 (2024)
4. Chen, H., Gomez, C., Huang, C.M., Unberath, M.: Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review. NPJ digital medicine **5**(1), 156 (2022)
5. Cho, S.M., Wu, W., Kilmer, E., Taylor, R.H., Unberath, M.: Feeling the stakes: Realism and ecological validity in user research for computer-assisted interventions. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 189–197. Springer (2025)
6. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

7. Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., Pinto dos Santos, D.: Automation bias in mammography: The impact of artificial intelligence bi-rads suggestions on reader performance. Radiology **307**(4), e222176 (2023)
8. Elendu, C., Amaechi, D.C., Okatta, A.U., Amaechi, E.C., Elendu, T.C., Ezeh, C.P., Elendu, I.D.: The impact of simulation-based training in medical education: A review. Medicine **103**(27), e38813 (2024)
9. Flake, J.K., Fried, E.I.: Measurement schmeasurement: Questionable measurement practices and how to avoid them. Advances in Methods and Practices in Psychological Science **3**(4), 456–465 (2020)
10. Gaube, S., Suresh, H., Raue, M., Lermer, E., Koch, T.K., Hudecek, M.F., Ackery, A.D., Grover, S.C., Coughlin, J.F., Frey, D., et al.: Non-task expert physicians benefit from correct explainable ai advice when reviewing x-rays. Scientific reports **13**(1), 1383 (2023)
11. Groh, M., Badri, O., Daneshjou, R., Koochek, A., Harris, C., Soenksen, L.R., Doraiswamy, P.M., Picard, R.: Deep learning-aided decision support for diagnosis of skin disease across skin tones. Nature Medicine pp. 1–11 (2024)
12. International Organization for Standardization: ISO 9241-210 Ergonomics of Human-System Interaction - Part 210: Human-Centred Design for Interactive Systems. https://www.iso.org (2010)
13. Jacobs, M., Pradier, M.F., McCoy Jr, T.H., Perlis, R.H., Doshi-Velez, F., Gajos, K.Z.: How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational psychiatry **11**(1), 108 (2021)
14. Jannin, P., Korb, W.: Assessment of image-guided interventions. In: Image-guided interventions: technology and applications, pp. 531–549. Springer (2008)
15. Lazar, J., Feng, J.H., Hochheiser, H.: Research methods in human-computer interaction. Morgan Kaufmann (2017)
16. Lekadir, K., Frangi, A.F., Porras, A.R., Glocker, B., Cintas, C., Langlotz, C.P., Weicken, E., Asselbergs, F.W., Prior, F., Collins, G.S., et al.: Future-ai: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. bmj **388** (2025)
17. Melles, M., Albayrak, A., Goossens, R.: Innovating health care: key characteristics of human-centered design. International Journal for Quality in Health Care **33**(Supplement_1), 37–44 (2021)
18. Park, Y., Jackson, G.P., Foreman, M.A., Gruen, D., Hu, J., Das, A.K.: Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA open **3**(3), 326–331 (2020)
19. Seah, J.C., Tang, C.H., Buchlak, Q.D., Holt, X.G., Wardman, J.B., Aimoldin, A., Esmaili, N., Ahmad, H., Pham, H., Lambert, J.F., et al.: Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. The Lancet Digital Health **3**(8) (2021)
20. Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., et al.: Human–computer collaboration for skin cancer recognition. Nature Medicine **26**(8), 1229–1234 (2020)
21. Vasey, B., Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., Denniston, A.K., Faes, L., Geerts, B., Ibrahim, M., et al.: Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: Decide-ai. bmj **377** (2022)

22. Vazire, S., Schiavone, S.R., Bottesini, J.G.: Credibility beyond replicability: Improving the four validities in psychological science. Current Directions in Psychological Science **31**(2), 162–168 (2022)
23. Wekenborg, M.K., Gilbert, S., Kather, J.N.: Examining human-ai interaction in real-world healthcare beyond the laboratory. NPJ Digital Medicine **8**(1), 169 (2025)
24. Yang, Q., Steinfeld, A., Rosé, C., Zimmerman, J.: Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In: Proceedings of the 2020 CHI conference on human factors in computing systems. pp. 1–13 (2020)
25. Yatani, K.: Effect sizes and power analysis in hci. In: Modern statistical methods for hci, pp. 87–110. Springer (2016)
26. You, F., Khakhar, R., Picht, T., Dobbelstein, D.: Vr simulation of novel hands-free interaction concepts for surgical robotic visualization systems. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 440–450. Springer (2020)