A Visual Analytics Framework for Assessing Interactive AI for Clinical Decision Support

Eric W. Prince† and Todd C. Hankinson

Department of Neurosurgery, University of Colorado Anschutz Medical Campus
Aurora, Colorado 80045, USA
†*E-mail: Eric.Prince@CUAnschutz.edu*

Carsten Görg

Department of Biostatistics and Informatics, Colorado School of Public Health
Aurora, Colorado 80045, USA

Human involvement remains critical in most instances of clinical decision-making. Recent advances in AI and machine learning opened the door for designing, implementing, and translating interactive AI systems to support clinicians in decision-making. Assessing the impact and implications of such systems on patient care and clinical workflows requires in-depth studies. Conducting evaluation studies of AI-supported interactive systems to support decision-making in clinical settings is challenging and time-consuming. These studies involve carefully collecting, analyzing, and interpreting quantitative and qualitative data to assess the performance of the underlying AI-supported system, its impact on the human decision-making process, and the implications for patient care. We have previously developed a toolkit for designing and implementing clinical AI software so that it can be subjected to an application-based evaluation. Here, we present a visual analytics framework for analyzing and interpreting the data collected during such an evaluation process. Our framework supports identifying subgroups of users and patients based on their characteristics, detecting outliers among them, and providing evidence to ensure adherence to regulatory guidelines. We used early-stage clinical AI regulatory guidelines to drive the system design, implemented multiple-factor analysis and hierarchical clustering as exemplary analysis tools, and provided interactive visualizations to explore and interpret results. We demonstrate the effectiveness of our framework through a case study to evaluate a prototype AI-based clinical decision-support system for diagnosing pediatric brain tumors.

Keywords: Clinical Decision Making; AI-Supported Interactive Decision Making; Evaluation Studies; Visual Analytics Framework.

1. Introduction

Artificial Intelligence (AI) can transform healthcare decision-making by quickly analyzing large amounts of data and improving diagnostic accuracy and patient outcomes.[1] However, ethical and legal implications, transparency of AI algorithms, and integration into existing workflows present challenges that require careful management.[1,2] Although AI has been increasingly used to support decision-making across various fields, more studies are needed to safely and

efficiently enhance human judgment and interpretation. Achievieng this goal requires the evaluation of AI systems on their algorithmic performance and their impact on humanistic aspects.

A comprehensive and systematic approach is needed to assess the impact of AI on decision making, particularly in high-risk settings such as healthcare.[3,4] For example, a recent study demonstrated that GPT-4V frequently presents flawed medical rationales in cases where it makes the correct final choices regarding the interpretation of radiologic imaging.[5] Examples of clinical experts' interactions with AI systems[6,7] reveal a gap in understanding AI's impact on humanistic aspects of clinical decision-making.

This gap extends to developing objective and precise techniques to evaluate AI technologies' safety and predictive precision.[5] Such evaluation techniques are still a bottleneck in the translational pipeline from a prototype tool to clinical deployment.[8] Our research highlights the substantial challenges related to implementing AI in high-risk decision-making scenarios within healthcare. We present robust and scalable exploratory analysis methods for evaluating AI systems and facilitating their broader acceptance and implementation in healthcare decision-making.

Monitoring clinical AI software effectively ensures performance, compliance, innovation, and better patient outcomes through data analysis and personalized medicine. Our proposed framework uses regulatory guidelines and statistical methods to assess system factors. Developing clinical AI software requires a structured framework: defining the problem, collecting and preparing data, developing and evaluating the model, and implementing and monitoring it. This comprehensive approach ensures that the software addresses specific clinical tasks, uses relevant data, integrates into the clinical workflow, and stays up-to-date.

We introduce a scalable framework implemented as an interactive software solution to analyze AI's impact in high-risk clinical decision-making scenarios. Its goals include identifying subgroups, detecting outliers, and supporting compliance with regulations. We integrate methodologies from multiple fields, including factor analysis, hierarchical clustering, adherence to regulatory guidelines, and interactive visualizations, to thoroughly analyze and enhance AI effectiveness in clinical decision-making. An end-to-end evaluation framework can enhance healthcare decision-making by improving AI's effectiveness, facilitating its implementation, and promoting adherence to regulatory guidelines, potentially leading to better patient outcomes. We demonstrate the utility and effectiveness of our framework through a case study assessing a prototype AI-based clinical decision-support system for the diagnosis of pediatric brain tumors.

2. Background

AI-assistance for Clinical Interpretation on Radiographic Images of CNS Tumors

As a use case in high-risk clinical decision-making, we look to AI support for diagnosing and managing central nervous system (CNS) tumors. In this context, experts use demographics, clinical presentation, imaging, and molecular information[9] for tumor diagnosis. AI systems can support efficient detection, diagnosis, staging, prognosis, and treatment planning of brain

tumors, among other applications.[9,10] These clincial decisions are only sometimes clear-cut and can require significant resource allocation. It is generally agreed upon that AI has ample room to support clinical decision-making in this context.[6,11,12] However, when considering the humanistic aspects of clinical AI support, it is becoming increasingly apparent that AI has a heterogeneous impact on human decision-makers.[6,11,12] Human experts may exhibit automation bias or neglect, where they overweight and underweight the AI prediction relative to their own, respectively.[6] Therefore, assessing the effect AI assistance has on decision-makers at the system level is essential. It is important to note that although AI has the potential to enhance clinical decision-making significantly, it also brings challenges that need to be addressed. These challenges include data-related issues, digital inequity gaps, bias, and the need for robust governance frameworks that balance safety and innovation.[10]

Consensus Statements and Guidelines for Clinical AI

In the healthcare sector, specific guidelines have been established to rigorously evaluate the clinical impact of AI, ensuring standards for transparency and ethical adherence. These guidelines contrast with those of other sectors, such as finance. Frameworks such as TRIPOD-AI[13] and CONSORT-AI[14] provide structured recommendations for preclinical and clinical AI trials; they emphasize standardized reporting and detailed intervention analysis. The DECIDE-AI[15] guidelines serve a critical role in bridging the preclinical and clinical AI trial phases.

DECIDE-AI targets early-stage clinical evaluations of AI-driven decision-support systems, emphasizing the importance of assessing clinical utility, safety, and ergonomic factors to prepare for broader clinical trials. Developed through international consensus involving experts from diverse areas, these guidelines are pivotal in ensuring that AI technologies are safely and effectively integrated into clinical practices. We used DECIDE-AI to drive the design of our framework, aligning our evaluation methods with best practices for early-stage AI assessment in healthcare.

Visual Analysis of Qualitative Data

The qualitative data analysis software landscape mainly features commercial products, with a notable deficit in advanced open-source options tailored for specialized fields such as clinical AI. Although feature-rich, commercial software like NVivo and ATLAS.ti are expensive and designed for broader use, making them less suitable for niche research areas with limited budgets and cases.

We introduce a new visual data analysis tool designed specifically for early-stage clinical AI evaluations to address this gap. It offers a cost-effective, scalable solution for clinical AI studies, enhancing user-centered evaluations and supporting the development of tailored clinical AI applications.

3. Analytical Objective, Experimental Data, Regulatory Guidelines, and Interface Design

We previously presented a framework for designing, implementing, and evaluating clinical AI tools from an implementation science perspective.[16] Here, we introduce a new framework for

the analysis phase of application-based studies. Specifically, we consider (a) how users can interact with AI systems to make sense of patient data so that they can make effective care decisions and (b) how we monitor these AI systems for safety and efficacy (Figure 1).

We emphasize an exploratory and holistic mindset when interpreting the results of AI evaluation studies. Our framework provides an overview of the data collected in the experiment, complemented by secondary views that can display various facets that detail aspects of the experimental data. We strive for simplicity and efficiency, integrating a minimalistic user interface and implementing linked-view mechanisms for seamless visual filtering. Below, we present the experimental data used to inform our design choices and describe the design of the primary and secondary views.

Following regulatory guidelines, such as DECIDE-AI, to lead analysis is essential when developing AI for clinical decision support. Figure 1 depicts some of these guidelines as black-and-white text boxes. This structured method improves system development, ensures healthcare compliance, and thoroughly evaluates AI integration. It is important for creating efficient, secure systems. Our framework supports identifying personas and patterns in evaluation data and aligns with the DECIDE-AI guidelines.

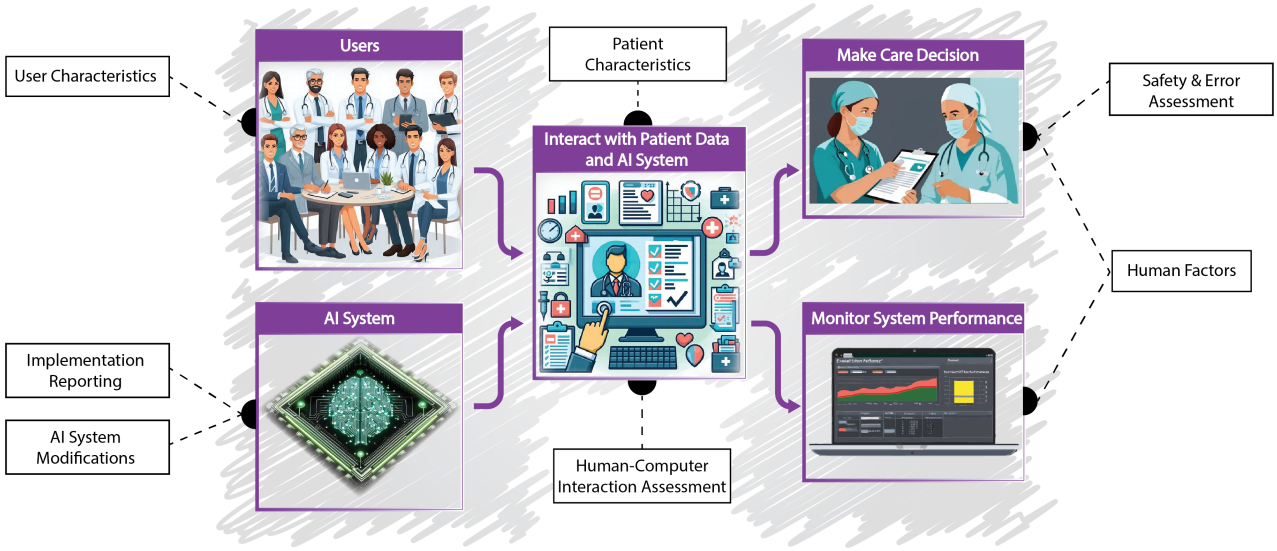

Fig. 1. Conceptual depiction of how users and interactive AI systems come together to make care decisions while monitoring system performance. DECIDE-AI themes guiding clinical AI evaluation are shown in black-and-white text boxes.

Analytical Objective: Identifying Personas and Patterns

Our analytical initiative is focused on coarse but comprehensive data exploration. This task plays a significant role in the initial phases of clinical AI system development to support identifying user personas and patient subgroups, as well as detecting patterns of human-AI agreement and disagreement. This exploration aims to guide the development of AI systems tailored to their users and environments, resulting in more personalized and relevant applica-

tions. During this stage, it is essential to acknowledge the subtle interactions between users and AI systems that can yield valuable insights into refining AI algorithms for optimal performance in real-world clinical settings.

Experimental Data

When evaluating AI's role in clinical decision-making, it is important to adopt a comprehensive approach that considers user-centered design and patient-related data. This approach is centered around the collection of a broad range of categorical and continuous variables describing both the performance of the AI system and the interactions of the user with the system and patient data.

In user-centered design, categorical data (nominal and ordinal) is essential for categorizing and understanding user interactions and experiences. Nominal data can reveal usage patterns and tool preferences, such as user roles (e.g., doctors, nurses, administrators) and AI tool types (e.g., diagnostic aid, treatment planner). Ordinal data, such as user satisfaction ratings or task difficulty levels, can provide insight into the usability and effectiveness of AI tools. Meanwhile, continuous data, including interval and ratio data, provide quantitative user engagement and tool performance measures. Interval data, such as response times or system up-time, and ratio data, such as usage counts, session lengths, or error rates, provide precise metrics to track changes over time or after modifications.

In addition to user-centered data, it is equally important to gather patient-related data. Categorical patient data, such as diagnosis (e.g., Central Nervous System (CNS) tumor), treatment type (e.g., surgery, radiation therapy, chemotherapy), and genetic markers, offer essential insights into the patient's health status and the complexity of their case. Similarly, continuous data points such as tumor size, biomarker levels, and treatment response (e.g., tumor size changes or patient symptoms over time) play a pivotal role in providing precise and quantifiable measures of the patient's condition and treatment progress.

Taking into account both user and patient data, the AI tool can be designed to provide a more holistic and personalized user experience. It can cater to the user's specific tasks, such as diagnosing a CNS tumor or monitoring a patient's response to treatment, thereby enhancing the tool's effectiveness and usability in the clinical setting. This comprehensive data collection and consideration approach is fundamental in the user-centered design and evaluation of AI tools in clinical settings. It ensures that the tool meets the user's needs and improves patient outcomes, which is the ultimate goal of healthcare delivery. Thus, collecting and considering diverse data types is fundamental in evaluating and optimizing AI in a clinical setting. It also aids in mitigating biases and improving the fairness and equity of AI-driven clinical decisions.

Themes of Regulatory Guidelines Driving the Design

To achieve our analytical objective, we lean on the themes and guidelines in the DECIDE-AI framework.[15] Each theme is tailored to glean critical insights during the early phases of AI system development and deployment in healthcare settings. The remainder of this subsection provides a summary of each of these themes.

User Characteristics Analysis. This theme involves collecting and assessing demographic and clinical data from healthcare providers to develop practical AI solutions that meet diverse user needs. This strategy enhances the system's versatility and facilitates its acceptance and integration in clinical settings. The theme aligns with DECIDE-AI guidelines 9A and 9B.

Implementation Reporting. This theme analyzes user interaction with the AI system and its impact on clinical workflows, focusing on user engagement and system acceptance. The goal is to ensure that the AI improves existing workflows and is easily integrated into clinical settings. This theme aligns with DECIDE-AI guidelines 10A and 10B.

AI System Modifications. To maintain the AI system's effectiveness and meet its users' needs, it is imperative to document all modifications made during the study and analyze their impact on the system's outcomes. This theme is essential for the system's continued evolution and clinical efficacy. It corresponds with DECIDE-AI guideline 11.

Human-Computer Interaction Assessment. This theme assesses user agreement and compliance with AI recommendations, focusing on improving trust and system reliability. By analyzing deviations, developers can refine the AI to better meet user expectations and ensure its recommendations are practical for integration into daily operations. This theme aligns with DECIDE-AI guideline 12.

Safety and Error Analysis. This theme focuses on identifying and addressing errors, malfunctions, potential risks, and observed harm in the AI system to safeguard patient safety. Vigilant monitoring and mitigation ensure compliance with healthcare regulations and ethical technology deployment in clinical settings. This theme aligns with DECIDE-AI guidelines 13A and 13B.

Human Factors Analysis. This theme combines usability testing and learning curve evaluations to ensure the AI system is practical and accessible from initial use to complete competence. Meeting practical needs and improving user experiences provides high user adoption and satisfaction and aligns with DECIDE-AI guidelines 14A and 14B.

Due to space constraints, we focus on the themes of Implementation Reporting, Human-Computer Interaction Assessment, and Human Factor Analysis in the remainder of this paper.

Interface Design

This section outlines our interface design, which follows a top-down conceptual approach. The UI is organized into primary and secondary views, creating a light, focused layout that enhances interaction. This hierarchical structure improves user efficiency by providing coarse overviews and allowing granular analysis of selected topics. Interactive views enable dynamic data filtering, enriching the user experience with structured navigation and focused content.

Design of Primary View. Figure 2 shows an example of the primary view, detailed in the case study in Section 4. Our visual analytics framework supports a two-stage, top-down analytical process for handling complex clinical datasets. The initial analysis uses a full-screen plotting window for broad pattern recognition and preliminary insights.

The primary view is configured with interactive functionalities to transition to the finer data exploration phase. These include dynamic linking capabilities between primary and secondary data views and providing contextual information via tooltips. Such features are in-

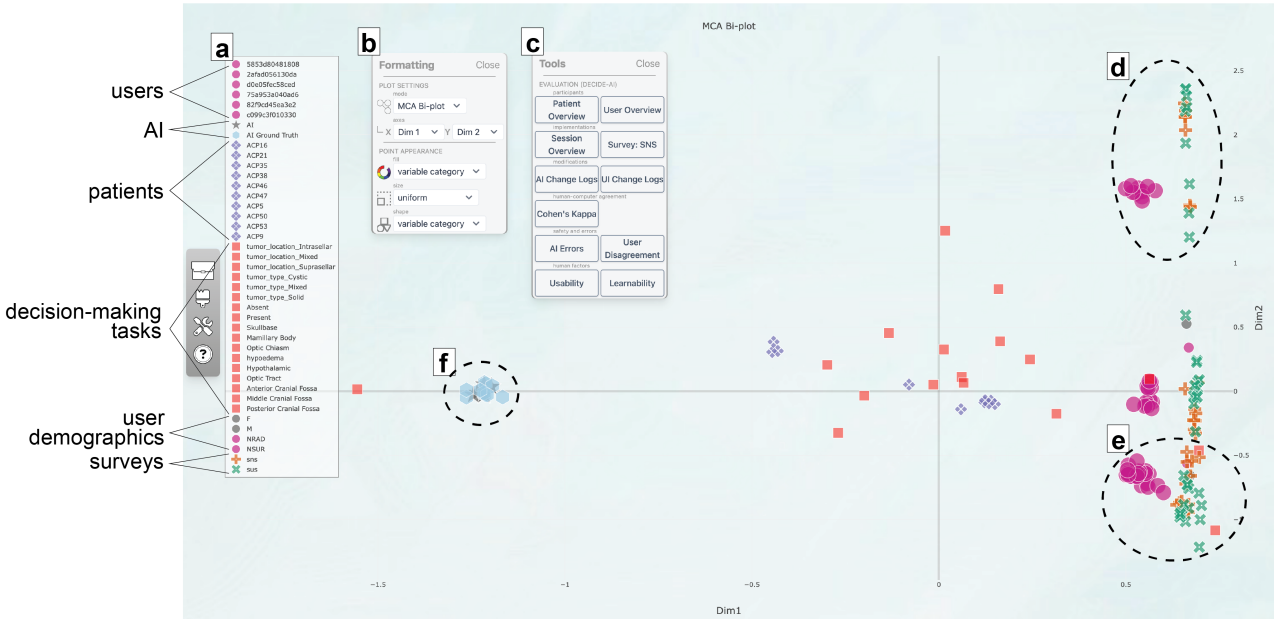

Fig. 2. Overview of the primary view displaying a factor analysis (i.e., MCA) bi-plot of clinical AI user study experimental data. (a) Legend for the plot depicting entities and variables within the study dataset. Entities and variables are annotated based on whether they are a user, AI agent, patient, decision-making task, demographic, or survey value. (b) The Formatting panel controls the display of the primary view and the appearance of marks. Marks are currently double-encoded for color and shape, showing variable categories. (c) The Tools panel contains buttons to toggle secondary views. The buttons are organized according to DECIDE-AI guidelines. (d, e) Examples of participants that represent 2 Personas. (f) Grouping of AI-predicted and AI-ground truth values.

dispensable for users focusing on detailed data inspections, where precision in isolating and scrutinizing data segments is necessary. Interactive tools like zoom, adjustable filters, and data point selection (rectangular or lasso) enhance query specificity, streamline workflows, and deepen analysis.

The minimalist toolbar of the interface, shown adjacent to Figure 2a on the left side, maintains simplicity by housing navigational buttons like project, formatting, tools, and help. This design choice preserves an intuitive navigation structure while supporting extensive functionality, minimizing cognitive load for the analyst.

For this example, our primary view layout is determined using multiple correspondence analysis (MCA). This type of factor analysis optimally suits the assessment of nominal categorical data like surveys. We expand on other factor analyses in our Discussion below. Using a unified graphical interface with dual-coding (glyphs and colors) helps understand the relationships between clinicians' behaviors, patient data, and AI insights.

The configuration of the primary view (Figure 2b), therefore, elevates the analytical capabilities required in clinical settings and aligns with rigorous academic data processing and visualization standards. Designing to meet users' operational and cognitive needs supports nuanced data exploration, which is essential for advancing AI in healthcare and evaluating its impact on clinical decision-making.

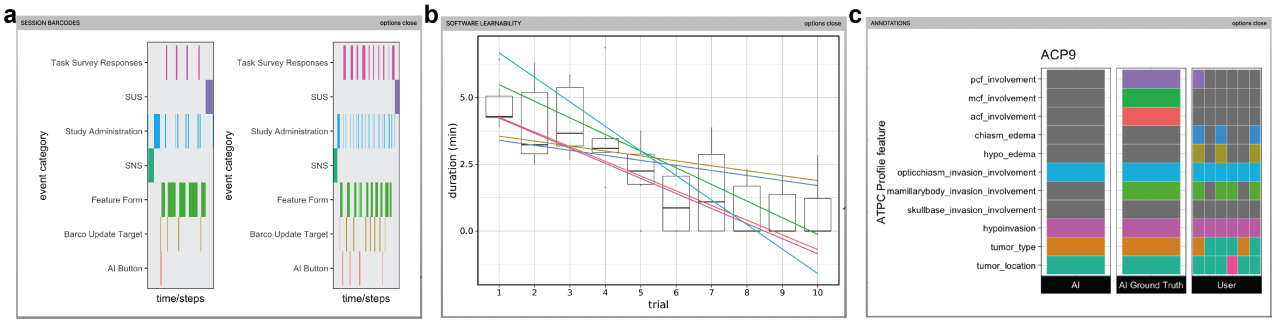

Fig. 3. Comparative Analysis of User Sessions and AI Interactions. (a) This panel illustrates interactions across two distinct sessions, capturing surveys, administrative actions (e.g., 'Next' button clicks), and specialized tasks (e.g., survey responses). The utility of AI features is examined through variations in image viewing (Barco Update Target) and AI button use, indicating differing reliance on AI tools between sessions. (b) This diagram shows a trend of decreasing task completion times, indicating improved user proficiency with system utilization over the session. (c) A heatmap highlights alignment and discrepancies in decision-making annotations between AI predictions, AI ground truth, and user selections. This visualization is instrumental in evaluating the AI's alignment with user decisions and overall influence on decision-making.

Design of Secondary Views. The secondary views in our visual analytics framework are intricately designed to complement the primary view by providing enhanced functionality for detailed and task-specific analysis, as illustrated in Figure 3. These views, which can be triggered from the toolbar menu (Figure 2c) have been developed with particular considerations to support the themes within DECIDE-AI required to effectively evaluate AI interactions within clinical contexts.

The decision to implement popup windows for secondary views is purposeful. It is designed to preserve the primary interface's clarity while enabling access to advanced data inspection when required. This approach allows users to engage with complex data sets without cluttering the primary view, facilitating user-controlled complexity in the visualization environment. Such a design is critical for tasks requiring focused analytical attention on specific data subsets while maintaining sight of the broader analytical context.

The secondary views utilize responsive SVG display widgets, which are pivotal for the dynamic visualization of intricate, multi-dimensional data typical in clinical analytics. These widgets are essential for detailed data relationship analyses, especially for interactions between patient data and AI outputs, as they allow users to interactively manipulate visual elements.

The ability to resize and reposition popup windows empowers users to tailor the analytical workspace to their specific needs or preferences, enhancing the ergonomics of data analysis. This flexibility is essential during analyses such as cross-referencing multiple data sources or adjusting visual layouts to better interpret data correlations and trends. Combining secondary views with the primary view allows for both broad and detailed questions to be addressed simultaneously.

4. Case Study

To demonstrate our framework, we used data from a previously conducted evaluation study of a prototype AI-based clinical decision-support system for the diagnosing pediatric brain tumors. For context, we summarize the system and study here; details are provided in our previous work.[17]

Development and Evaluation Study of Interactive AI Clinical Decision Support Software

User Interface and Radiology Workstation Simulator. We collaborated with clinical partners in a step-by-step design study, collecting visualization samples, conducting user interviews, and improving designs based on feedback. In the initial phase, we used visualizations to display the performance of the AI model. In the second phase, we visualized predictions for clinicians using existing tools and conducted a user study. After immersing ourselves in the clinical environment, we refined our task specifications and created initial prototypes in the third phase. Finally, we built a radiology reading terminal and implemented basic AI interfaces as web applications in the fourth phase. Additional details are provided in our previous publication.[18]

AI Model Backend. The study used the ATPC50 dataset from the Advancing Treatment for Pediatric Craniopharyngioma (ATPC) international multi-institutional consortium in North America, which included information from 50 ACP patients.[17] The study focused on patients' initial presentations, utilizing imaging data from preoperative CT and MRI scans, with radiographic features annotated by a certified neuroradiologist. The AI model thoroughly preprocessed DICOM inputs by resizing images, adjusting contrast, and simulating different patient positions. The data was rescaled to the JPEG range and then processed using ResNet V2 techniques. The study also included using a variational autoencoder for data reconstruction and deep learning classifiers for diagnostic analysis.[17,19]

Experimental Study Design. The study recruited six post-residency faculty attending clinicians (three females and three males) from Children's Hospital Colorado, focusing on those specializing in neurosurgery and neuroradiology. Participants were recruited via email and scheduled for individual 30-minute sessions over a two-week period to accommodate their busy schedules.

At the start of each session, participants shared demographic information, were introduced to the study's goals, and completed the Subjective Numeracy Scale (SNS) survey. They then received a step-by-step guide to the AI decision support tool through ten instructional slides. Participants used radiologic images of CNS tumors to annotate an 11-point feature profile of a pediatric CNS tumor known as Adamantinomatous Craniopharyngioma, both with and without AI support. These feature profiles were completed within the software as a form with checkboxes.

Participants engaged with AI in two forms. The first was a passive AI assistant that flagged a checkbox if the user selected a value that was different from the AI prediction. The second was a direct AI assistant that provided users with the AI-predicted feature profile and a list of other patients that the AI model suggested were similar, based on L1 distance between

prediction vectors. At the end of the session, participants provided feedback by completing the System Usability Scale (SUS) survey. We collected survey data, feature predictions, and system interaction logs.

Exploration and Interpretation of Data from the Evaluation Study

We now describe how our new framework for assessing interactive AI for clinical decision support was used to analyze the data collected in our evaluation study.

Our analysis, which encompassed survey responses and system interaction logs, distinguished two primary user personas: the 'Tech Novice Numerate' (Figure 2d) and the 'Confident Numerate' (Figure 2e). The 'Tech Novice Numerate' users displayed moderate numerical skills but struggled to navigate the AI system, indicating a pressing need for improvements in interface design and enhanced user training. In contrast, the 'Confident Numerate' users, who demonstrated high numerical proficiency, expressed concerns about the system's consistency, suggesting potential reliability and user acceptance issues.

An in-depth examination of the utilization of AI tools revealed significant variances in the degree of dependency on AI support, as observed through differential usage of the "Barco Update" feature for additional image views and "AI button" interactions. Additionally, a chronological analysis of task completions, encompassing SUS, SNS, and Feature Form responses, shed light on the users' learning trajectories and the system's adaptability throughout the session.

In Figure 2f, the overlay of points for AI predicted values and the AI ground truth suggests a high degree of agreement between the AI model's predictions and the annotations made by a board-certified clinical expert, considered the 'ground truth' in this context. This expert is highly skilled and certified in the task at hand within this specific use case.

The fact that the AI model aligns closely with the ground truth annotator indicates that the model has learned to mimic the decision-making process of this particular expert quite accurately. However, it is essential to note that this expert may have interpretations that differ from other experts in the field. This is a common occurrence in many professional fields, including clinical practice, where different experts may have slightly different interpretations or approaches based on their training, experience, and personal biases.

Collaboration among human experts in clinical practice is crucial. Discussing interpretations with colleagues can help reach a consensus or understand different viewpoints, which can help mitigate discrepancies between different human experts. This collaborative approach is particularly important in the context of the figures, as it can help reconcile differences between the annotations that fall into Figure 2f (where the AI and the ground truth annotator agree) versus those in Figure 2d or 2e (where there may be disagreement).

Understanding potential biases in the AI model is essential for evaluating clinical AI devices in real-world settings. If the AI model consistently aligns with one expert (the ground truth annotator, in this case), it may indicate that the model is biased towards that individual's interpretations. These models need to generalize well across different experts and not just mimic the decisions of one individual.

This granular analysis of user-system engagement deepens our understanding of behavioral dynamics and provides actionable insights for targeted enhancements in AI system design and interface. These empirical findings emphasize the critical role of user-centered design in developing intuitive and reliable clinical decision-support tools, enhancing system functionality, and fostering greater user trust and satisfaction in healthcare AI applications.

5. Discussion and Lessons Learned

Integrating AI in decision-making across high-stake sectors underscores a transformative shift towards data-driven practices. However, deploying these AI systems, particularly in sensitive areas such as healthcare care, requires an approach that couples algorithmic insights with indispensable human judgment. The predominant reliance on commercial products often leaves gaps in affordability and customization, especially in specialized fields such as clinical AI. By introducing a visual analytics framework purposely built for clinical AI applications, we propose a solution tailored to meet these unique requirements. Our framework can advance analysis capabilities by interpreting data from clinical user studies and increasing accessibility and practical relevance, reducing dependency on costly and often overly complex tools.

Consideration of the humanistic aspects of clinical AI evaluation is essential for several reasons. Real-world scenarios vary significantly from controlled experiments, making evaluating AI tools with diverse patient populations and varying data quality across discrete clinical tasks to ensure their generalizability. Evaluations help to identify and mitigate biases inherent in healthcare systems, ensuring fairness and equity. Real-world testing is vital in revealing potential safety issues and unintended consequences, guaranteeing that AI tools perform accurately and reliably in clinical settings. Furthermore, realistic evaluations consider how AI integrates into existing workflows, including integration challenges, user experience, and impact on efficiency. Involving clinicians and patients in the evaluation process provides valuable insights into user acceptance, trust, and willingness to adopt AI tools, informing necessary improvements. Finally, adhering to regulatory guidelines, such as DECIDE-AI, significantly enhances the robustness and generalizability of clinical AI tools by emphasizing fundamental principles. These include risk assessment and benefit analysis in real-world contexts, encouraging external validation and independent testing, assessing clinical utility, and promoting transparency through clear documentation.

Basing a clinical AI evaluation method on factor analysis can enhance scalability and accommodate diverse data types. Evaluative efforts for clinical AI systems can generate a large volume of multiple data types. Empirical tools used in this field often involve survey methods that can gather character descriptions of users (e.g., demographics), information about system usability and a way to measure how well users can complete the specific task supported by the system. In addition, continuous numeric data is also relevant in this space with aspects like predictive probabilities from the AI model, system response time, user interaction metrics, and human error rates. Understanding and considering all aspects of the evaluation, including patient data, human expert judgment, and AI software interactions is important. This comprehensive understanding is what produces robust tools that fundamentally improve patient care.

Factor analysis is a method for identifying latent factors, or underlying variables, in observed data. Factor analysis uses the correlation structure amongst observed variables to model fewer unobserved, latent variables known as factors. Researchers use this statistical method when subject-area knowledge suggests that latent factors cause observable variables to covary. For instance, we can evaluate an expert's diagnostic prediction using patient data accessed in software and compare it to the validated diagnosis. The prediction may need to be corrected due to unobservable software interaction patterns, which are observable in factor analysis. By capturing shared variance, it simplifies complex relationships among variables, aiding in the simplification of data analysis. This method is also scalable, enabling the efficient handling of large datasets by reducing dimensionality and making computations more manageable.

Factor analysis can accommodate diverse data types, including continuous and categorical data, allowing for incorporating survey responses (categorical) and continuous data into factor models. For example, Multiple Factor Analysis (MFA) is a multivariate method used to study tables where a group of individuals is described by a set of variables, which can be quantitative and qualitative and are structured in groups. It is an extension of Principal Component Analysis (PCA) for quantitative variables, Multiple Correspondence Analysis for qualitative variables, and Factor Analysis of Mixed Data for variables that belong to both types.

We implemented factor analysis using MCA for our framework because the data from our evaluation study were mainly qualitative. However, many other factor analysis methods are available, such as nonlinear PCA, which handles mixed data types more effectively.[20] The selection of the factor analysis method is flexible, and we will explore this area further in future work to identify more sophisticated representations of this complex experimental context.

Effectively evaluating AI requires a delicate balance between realism and controlled experiments to ensure robustness and practical applicability. Multiple facets are involved in ensuring robust clinical AI software. One approach starts with simulated environments to understand fundamental behavior in controlled settings, allowing for controlled variation while maintaining reproducibility and gradually transitioning to real-world data. Standardized benchmark datasets can provide a baseline for performance comparison in controlled experiments, although it is essential to recognize their limitations in representing real-world complexity. Another valuable strategy is transfer learning, which entails training models on controlled data and fine-tuning them on real-world data to bridge the gap between controlled and realistic contexts. Field studies conducted in clinical settings with actual users are essential for observing how AI tools impact workflows, patient outcomes, and user satisfaction. Adversarial testing is also important, introducing realistic challenges such as noisy data and adversarial attacks during controlled experiments to reveal vulnerabilities and test robustness. When used collectively, these strategies contribute to a comprehensive and balanced approach to AI evaluation. This approach ensures that all aspects of AI performance are thoroughly tested and evaluated, providing a fair and thorough assessment of the system's capabilities.

An example of the need to consider the reality of deployment in contrast with controlled experiments and statistical analysis can be seen in our study. Factor analysis is useful for evaluating user studies, especially with structured questionnaires and surveys. It identifies relationships between variables, simplifies data, and highlights key factors influencing responses.

This helps researchers understand patterns in feedback and make informed decisions about tool design and functionality. However, it mainly focuses on statistical relationships and might miss nuances in user interactions. For example, in our study, all participants consistently experienced passive AI, but active AI was less used, likely due to the flawed concept requiring comparisons without prior knowledge. This added complexity and confusion, which would not be evident through factor analysis alone. To address such issues, deeper qualitative investigations are necessary. These can include user interviews, observational studies, and detailed feedback sessions to understand the context and reasons behind user behaviors. This approach provides richer insights beyond statistical analysis, ensuring that AI tools are usable and practical in clinical settings. Combining quantitative and qualitative methods can lead to a more comprehensive evaluation and refinement of AI support systems.

In conclusion, while these strategies have advantages and potential challenges, they all play an important role in ensuring the practical evaluation of clinical AI tools. By proactively considering these points and addressing potential critiques, we can work towards more robust, ethical, and effective AI in healthcare.

References

1. S. A. Alowais, S. S. Alghamdi, N. Alsuhebany, T. Alqahtani, A. I. Alshaya, S. N. Almohareb, A. Aldairem, M. Alrashed, K. Bin Saleh, H. A. Badreldin *et al.*, Revolutionizing healthcare: the role of artificial intelligence in clinical practice, *BMC medical education* **23**, p. 689 (2023).
2. R. Khera, A. J. Butte, M. Berkwits, Y. Hswen, A. Flanagin, H. Park, G. Curfman and K. Bibbins-Domingo, Ai in medicine—jama's focus on clinical outcomes, patient-centered care, quality, and equity, *Jama* (2023).
3. C. A. Longhurst, K. Singh, A. Chopra, A. Atreja and J. S. Brownstein, A call for artificial intelligence implementation science centers to evaluate clinical effectiveness (2024).
4. Y. Park, G. P. Jackson, M. A. Foreman, D. Gruen, J. Hu and A. K. Das, Evaluating artificial intelligence in medicine: phases of clinical research, *JAMIA open* **3**, 326 (2020).
5. Q. Jin, F. Chen, Y. Zhou, Z. Xu, J. M. Cheung, R. Chen, R. M. Summers, J. F. Rousseau, P. Ni, M. J. Landsman, S. L. Baxter, S. J. Al'Aref, Y. Li, A. Chen, J. A. Brejt, M. F. Chiang, Y. Peng and Z. Lu, Hidden flaws behind expert-level accuracy of multimodal gpt-4 vision in medicine, *npj Digital Medicine* **7**, p. 190 (Jul 2024).
6. N. Agarwal, A. Moehring, P. Rajpurkar and T. Salz, *Combining human expertise with artificial intelligence: Experimental evidence from radiology*, tech. rep., National Bureau of Economic Research (2023).
7. F. M. Calisto, J. Fernandes, M. Morais, C. Santiago, J. M. Abrantes, N. Nunes and J. C. Nascimento, Assertiveness-based agent communication for a personalized medicine on medical imaging diagnosis, in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023.
8. S. Reddy, W. Rogers, V.-P. Makinen, E. Coiera, P. Brown, M. Wenzel, E. Weicken, S. Ansari, P. Mathur, A. Casey *et al.*, Evaluation framework to guide implementation of ai systems into healthcare settings, *BMJ health & care informatics* **28** (2021).
9. B. H. Kann, A. Hosny and H. J. Aerts, Artificial intelligence for clinical oncology, *Cancer Cell* **39**, 916 (2021).
10. S. Khalighi, K. Reddy, A. Midya, K. B. Pandav, A. Madabhushi and M. Abedalthagafi, Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment, *NPJ Precision Oncology* **8**, p. 80 (2024).

11. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in *Proceedings of the AAAI conference on artificial intelligence*, (01)2019.
12. D. C. Chan, M. Gentzkow and C. Yu, Selection with variation in diagnostic skill: Evidence from radiologists, *The Quarterly Journal of Economics* **137**, 729 (2022).
13. G. S. Collins, P. Dhiman, C. L. A. Navarro, J. Ma, L. Hooft, J. B. Reitsma, P. Logullo, A. L. Beam, L. Peng, B. Van Calster *et al.*, Protocol for development of a reporting guideline (tripod-ai) and risk of bias tool (probast-ai) for diagnostic and prognostic prediction model studies based on artificial intelligence, *BMJ open* **11**, p. e048008 (2021).
14. X. Liu, S. C. Rivera, D. Moher, M. J. Calvert, A. K. Denniston, H. Ashrafian, A. L. Beam, A.-W. Chan, G. S. Collins, A. D. J. Deeks *et al.*, Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension, *The Lancet Digital Health* **2**, e537 (2020).
15. B. Vasey, M. Nagendran, B. Campbell, D. A. Clifton, G. S. Collins, S. Denaxas, A. K. Denniston, L. Faes, B. Geerts, M. Ibrahim *et al.*, Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: Decide-ai, *bmj* **377** (2022).
16. E. Prince, T. C. Hankinson and C. Görg, Easl: A framework for designing, implementing, and evaluating ml solutions in clinical healthcare settings, in *Machine Learning for Healthcare Conference*, 2023.
17. E. W. Prince, D. M. Mirsky, T. C. Hankinson and C. Görg, Impact of ai decision support on clinical experts' radiographic interpretation of adamantinomatous craniopharyngioma, in *AMIA Annual Symposium Proceedings*, 2024 (to be released November 2024).
18. E. W. Prince, T. C. Hankinson and C. Görg, The iterative design process of an explainable ai application for non-invasive diagnosis of cns tumors: A user-centered approach, in *2023 Workshop on Visual Analytics in Healthcare (VAHC)*, 2023.
19. E. W. Prince, R. Whelan, D. M. Mirsky, N. Stence, S. Staulcup, P. Klimo, R. C. Anderson, T. N. Niazi, G. Grant, M. Souweidane *et al.*, Robust deep learning classification of adamantinomatous craniopharyngioma from limited preoperative radiographic images, *Scientific reports* **10**, p. 16885 (2020).
20. M. Linting and A. Van der Kooij, Nonlinear principal components analysis with catpca: a tutorial, *Journal of personality assessment* **94**, 12 (2012).