AI-based Visual Support for Clinical Diagnosis of Pediatric Suprasellar Tumors and Impacts on Decision-Making Confidence

University of Colorado Anschutz Medical Campus

Eric W. Prince
University of Colorado

David M. Mirsky
Children's Hospital Colorado

Todd C. Hankinson
Children's Hospital Colorado

Carsten Görg
Colorado School of Public Health

The Children's Hospital

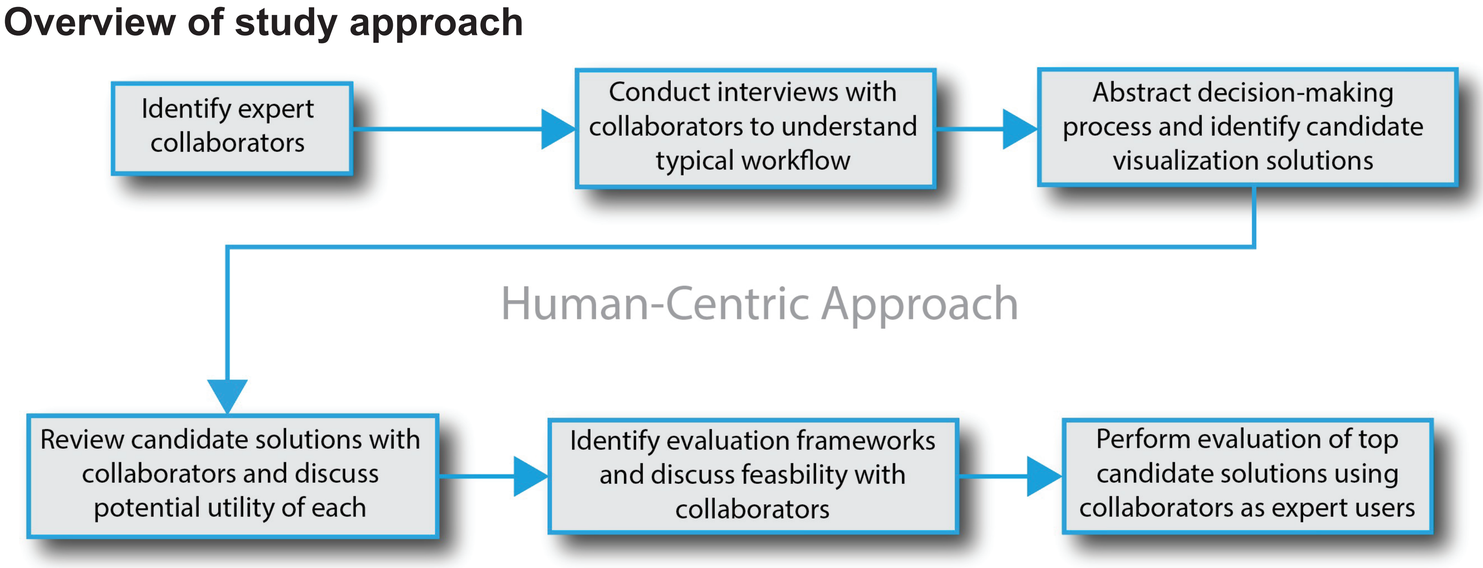
Background – Explaining AI-based predictions is fundamental for the development of clinical decision support systems.

A common visual approach for explaining imaging data predictions is to overlay saliency maps onto images to allow users to interpret what visual features are associated with a given prediction.

This approach can be difficult to utilize when differentiating nuanced concepts. For example, clinicians in neuro-oncology will commonly have to differentiate between a group of similar brain tumors (i.e., a radiographic differential diagnosis).

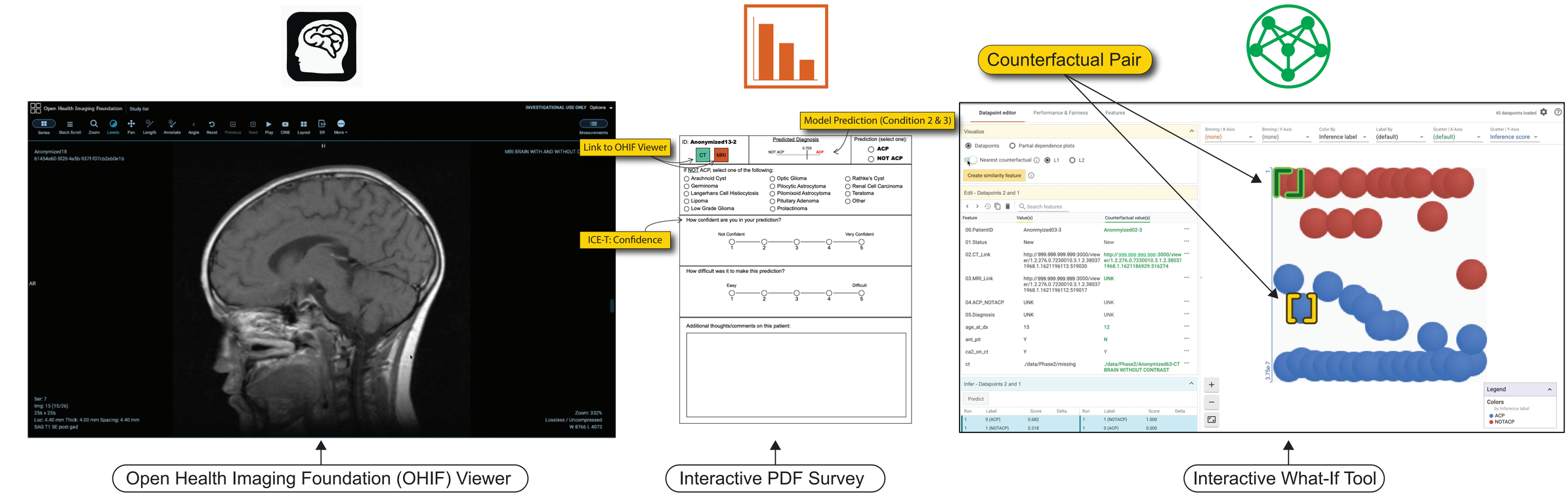
Hypothesis: visual representations of counterfactual conditions will enable clinicians to make diagnostic predictions more confidently, with less difficulty, and with greater accuracy.

Overview of study approach



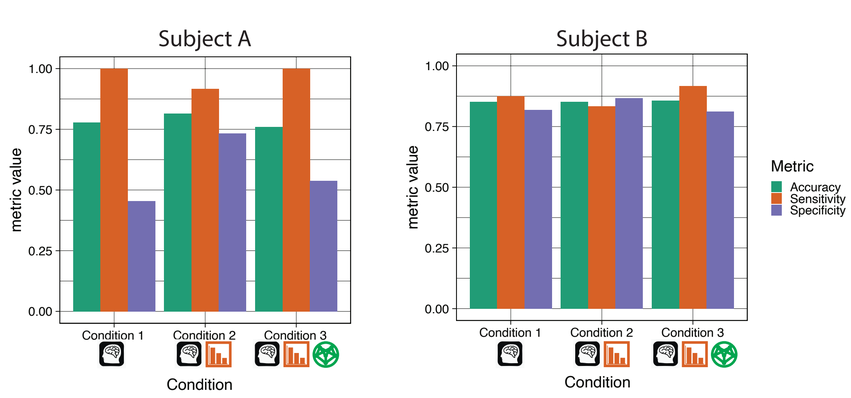
We met with a board-certified neurosurgeon and neuroradiologist to conduct interviews regarding their workflow for diagnosing suprasellar tumors. We surveyed the literature to derive a library of visualization methods and discussed which techniques would be most relevant. We then identified and adapted a survey-based evaluation metric (ICE-T) to assess how the visualization methods impact the confidence and perceived difficulty of a diagnosis for a clinician.

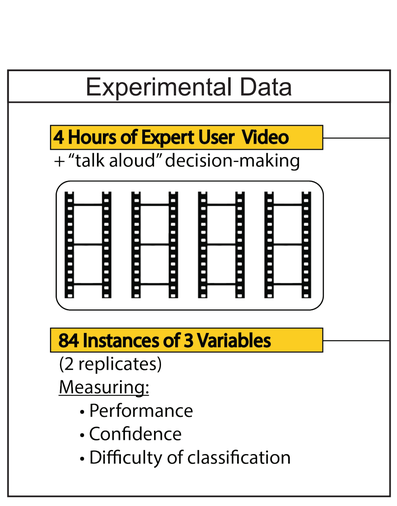
Experimental conditions

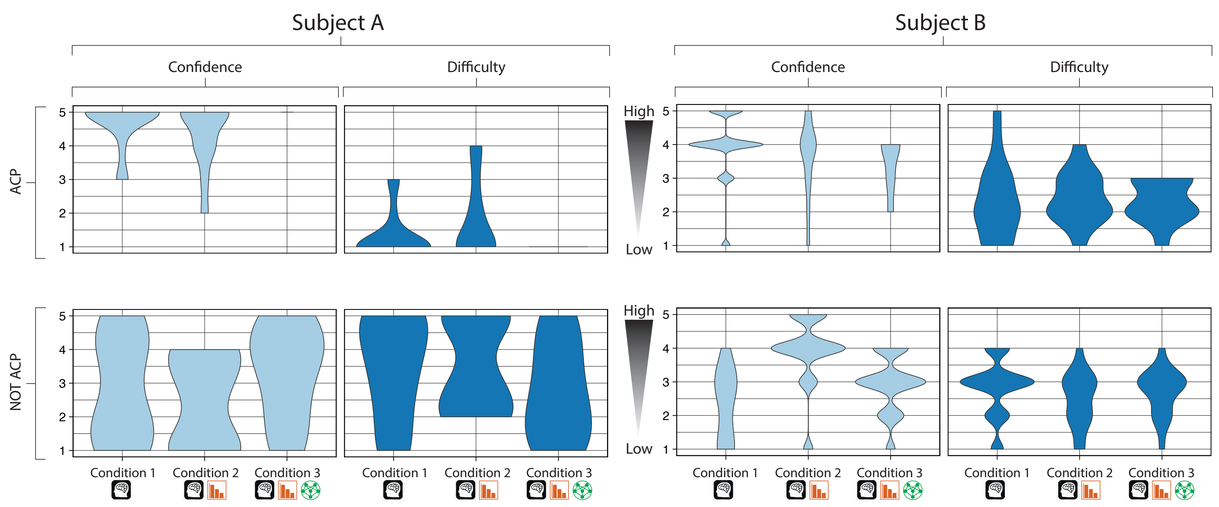


For all study conditions, subjects were given a set (n=28) of interactive PDF documents (center) which linked to the Open Health Imaging Foundation (OHIF) Viewer (left). The OHIF viewer provides a standard radiographic interpretation framework. Subjects were tasked with binary diagnostic prediction of adamantinomatous craniopharyngioma (ACP) versus other suprasellar tumors (NOTACP). The first condition provided only the OHIF Viewer. The second condition (center top) extends the first condition with a predicted value. The third condition extends the second condition with Google's What-If Tool (WIT; right). Subjects were prompted to think aloud during each condition and audio and screen recordings were captured for each session.

Results – There was no effect on the diagnostic performance of clinical users (right). There was no significant change in decision-making confidence and difficulty for each subject across the three study conditions for the NOTACP class of data. However, there was a trend for increased diagnostic confidence and decreased diagnostic difficulty for both subjects with predictions for the ACP class of data. This trend was strongest for the third condition of the study (OHIF Viewer + AI Prediction + WIT; bottom right).







Discussion – Both subjects utilized our intended functionality of the WIT which was to ask: "What are the most similar and dissimilar previously seen patients?".

In the course of querying this information using the WIT, subjects positively remarked that the counterfactual reasoning was very medically reasonable.

Unfortunately, the WIT was very difficult for users to engage with; a significant amount of time was required to explain the concept of the WIT as well as the software interface.

Our preliminary study has multiple limitations associated with the binary classification task not accurately reflecting a real clinical diagnosis and the use of a small single institutional dataset.

Future work will explore these limitations by performing similar studies with a multi-class classifier and a larger subject group who have not previously seen the clinical images.