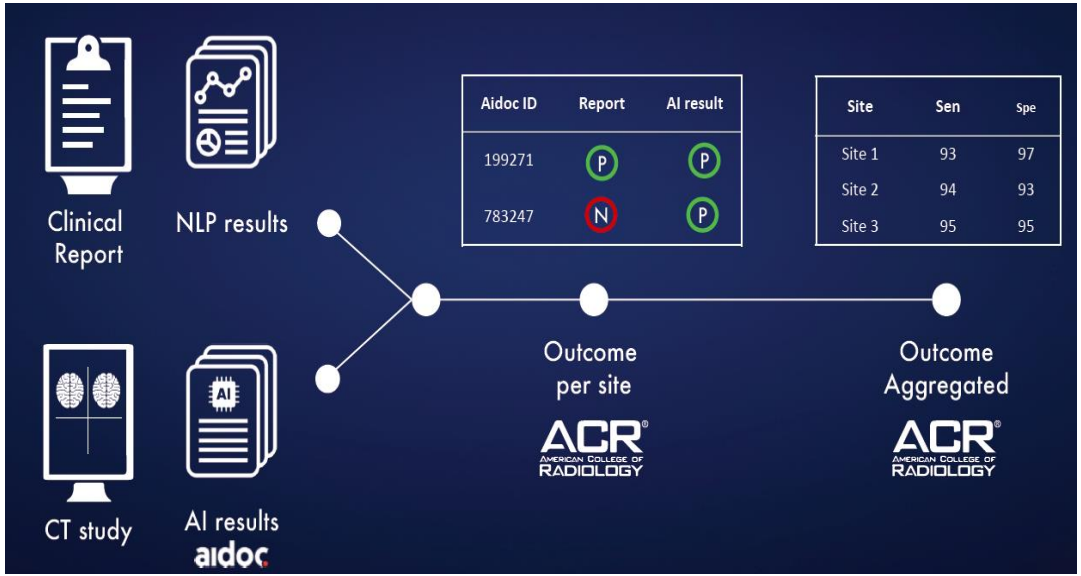
DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

Assess-AI and AI-PROBE: Monitoring Algorithm Performance in Clinical Practice

Axel W. E. Wismüller, M.D., M.Sc., Ph.D.[1,2]
[1]Department of Imaging Sciences
University of Rochester Medical Center, New York, U.S.A.
[2]Faculty of Medicine, Ludwig Maximilian University, Munich, Germany

ACR Assess-AI: How it works



- Workflow architecture for data submission to the ACR Assess-AI repository for post-market surveillance of a vendor-specific AI solution (Aidoc, Tel Aviv, Israel).
- URMC was first institution in the US that went live submitting data to the ACR Assess-AI repository (January 2020)
- AI reading results are compared with NLP results of final radiology reports, serving as case-specific ground truth.
- Aggregated information can be used by ACR to monitor diagnostic accuracy of such AI solutions.

DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

Quality Evaluation of AI in Clinical Practice

We start from the following assumption:

"A radiologist *with* AI will be better than a radiologist *without* AI."
(Keith Dryer)

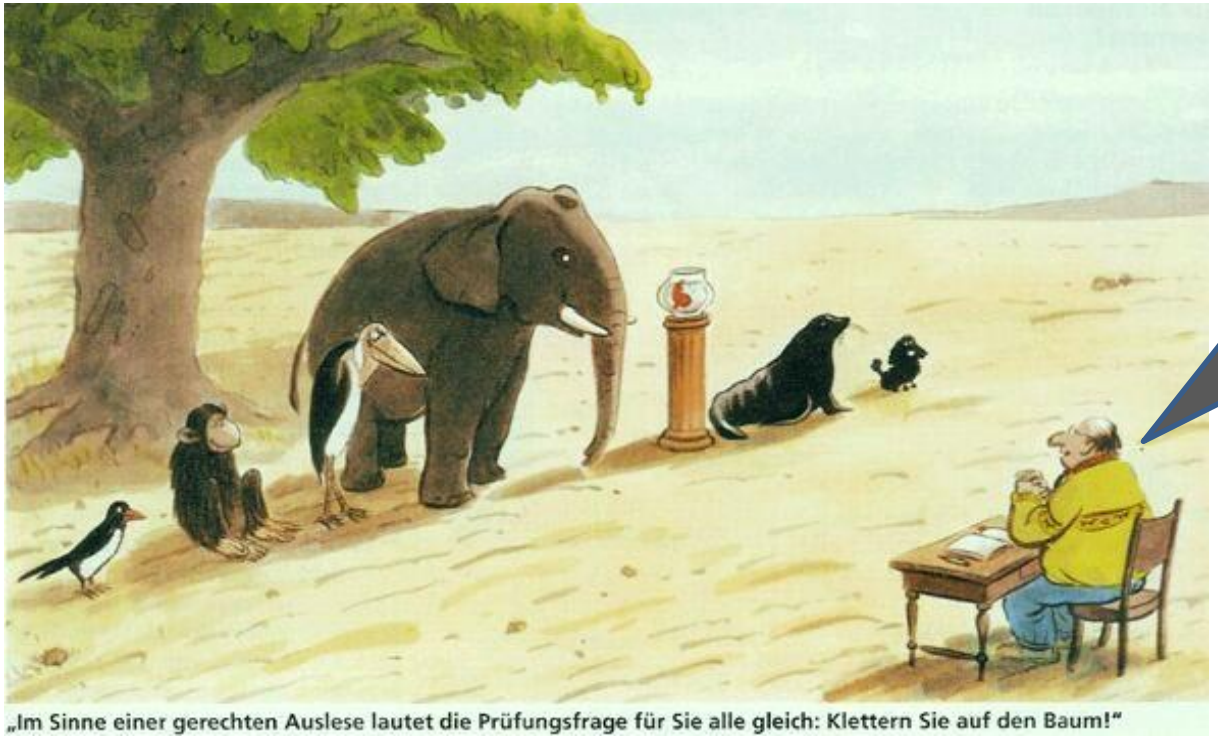
However: How can we test this hypothesis?

What determines the "goodness" of a radiologist *with* or *without* AI?

"Measure what is measurable, and make measurable what is not so."
(Galileo Galilei)

Challenge: We need to make the improvement caused by AI measurable.

Quality Evaluation of AI in Clinical Practice



„Im Sinne einer gerechten Auslese lautet die Prüfungsfrage für Sie alle gleich: Klettern Sie auf den Baum!"

To ensure a fair selection, the examination task *is the same* for all of you:

Climb the tree!

DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

Quality Evaluation of AI in Clinical Practice

- There is an urgent need for quantitatively evaluating the real-world practical usefulness of AI solutions in radiology.

- We need to perform this evaluation with the same rigor as evaluating a new drug for patient use.

Quality Evaluation of AI in Clinical Practice

Proposed solution:

Artificial Intelligence Prospective Randomized Observer Blinding Environment (AI-PROBE)

- Scientific approach for quantitative clinical performance evaluation of radiology AI systems within prospective randomized clinical trials.
- Our evaluation workflow encompasses a study design and a corresponding radiology Information Technology (IT) infrastructure that randomly blinds radiologists with regards to the presence of positive reads as provided by AI-based image analysis systems.

Quality Evaluation of AI in Clinical Practice

Application Example:

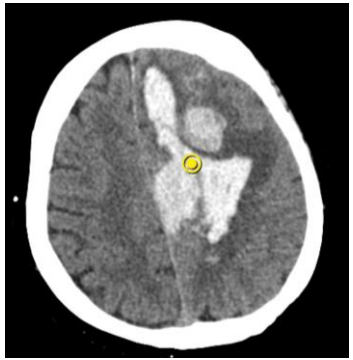
To demonstrate the applicability of our AI-evaluation framework, we present a first prospective randomized clinical trial on investigating the effect of automatic identification of Intra-Cranial Hemorrhage (ICH) in emergent care head CT scans on radiology study Turn-Around Time (TAT) in a clinical environment.

DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

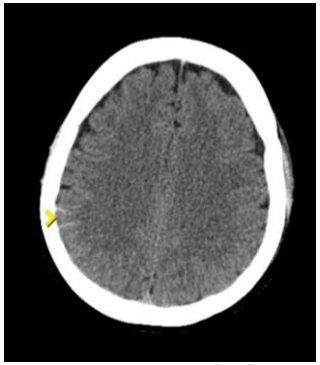
Methods: Imaging Data

- A total of 620 consecutive non-contrast head CT scans from two CT scanners used for inpatient and emergency room patients at a large academic hospital (University of Rochester Medical Center) were analyzed in this study.
- CT scans were prospectively acquired over a time period of 14 consecutive days.
- Immediately following image acquisition, scans were automatically analyzed for the presence of intracranial hemorrhage (ICH) using commercially available software (Aidoc, Tel Aviv, Israel).

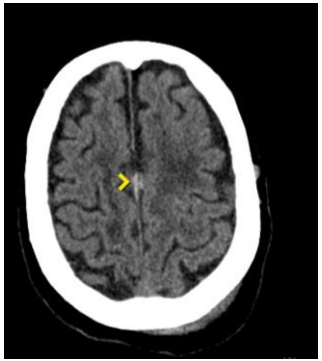
Examples: AI-based ICH Detection



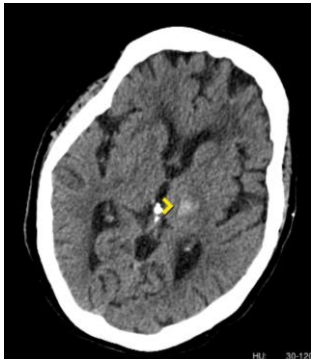
True positive



True positive



True positive

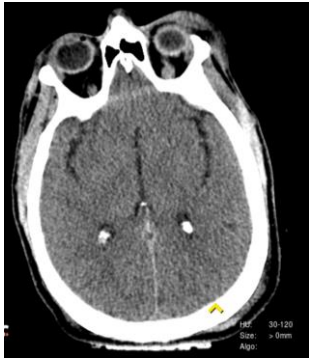


True positive

True positive

False positive

False positive



False positive

Methods: Study Design

- Cases identified as positive for Intracranial Hemorrhage (ICH) by AI (ICH-AI+) were automatically flagged in the radiologists' reading worklists, where flagging was randomly switched off with a probability of 50%.
- Study turnaround time (TAT) was measured automatically as the time difference between study completion time (=study accessible to radiologists for reporting) to study reporting time (=first report visible to clinicians, regardless whether preliminary or final).
- Time stamps for calculating TAT were automatically retrieved from various radiology IT systems.

DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

Methods: Statistical Analysis

- Turnaround times for flagged and non-flagged ICH-AI+ cases were compared using a one-sided *t*-test
- Diagnostic accuracy for all analyzed 620 CT studies was evaluated by calculating sensitivity, specificity, and accuracy for Intracranial Hemorrhage (ICH) detection.
- For this purpose, findings reported in the final radiology reports served as ground truth.

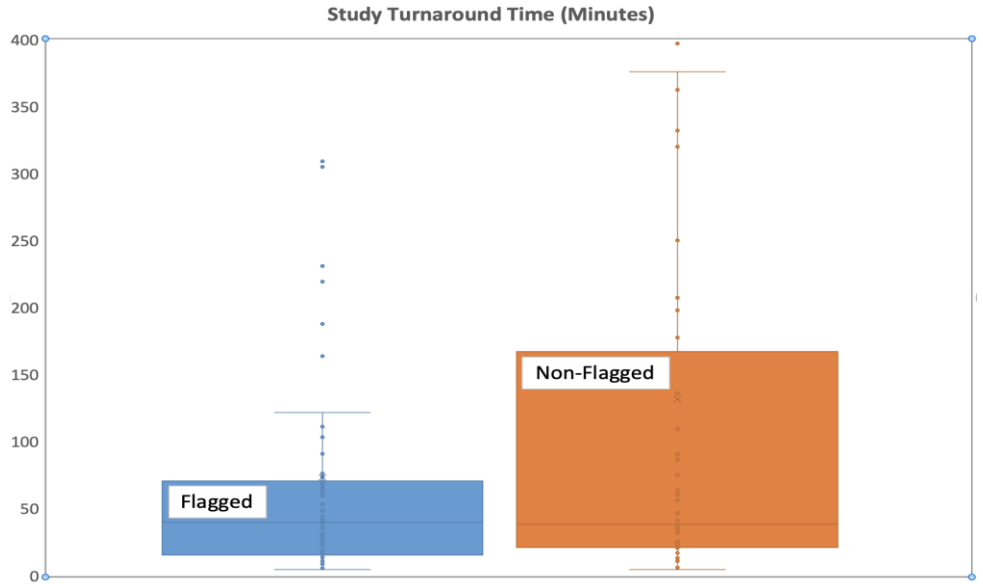
Results: Diagnostic Accuracy

- A total of 122 ICH-AI+ cases were found, of which 66 cases were flagged.

- 105 (85.9%) of the 122 ICH-AI+ cases were true positive reads.

- Diagnostic accuracy measures over all analyzed 620 cases:
  - Sensitivity: 95.0%
  - Specificity: 96.7%
  - Accuracy: 96.4%

Results: Turnaround Time



- Mean turnaround time for flagged cases (73 ± 143 min) was lower than mean turnaround time for non-flagged (132 ± 193 min) cases.
- Differences in turnaround time distributions for flagged and non-flagged cases were statistically significant ($p<0.05$, one-sided t-test).

Conclusion 1

- Automatic identification of intracranial hemorrhage provides high diagnostic accuracy, as indicated by our high sensitivity, specificity, and accuracy results.
- At first glance, this finding should carry the potential for improving clinical management by accelerating clinically indicated therapeutic interventions.
- To quantitatively evaluate this potential, we performed a prospective randomized clinical trial on the observable effect of automatic intracranial hemorrhage detection on turnaround times in emergency setting head CT scans.

DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

Conclusion 2

- Notifying radiologists on automatically detected ICH reduces turnaround times for reporting intracranial hemorrhage to clinicians in emergency setting head CT scans, as shown by our prospective, randomized clinical trial.

- Turnaround time reduction benefits for AI-based intracranial hemorrhage detection are likely to be even higher in outpatient setting head CT scans, where overall turnaround times are higher, and significant turnaround time differences have been reported by others*.

* See e.g.: M.R. Arbabshirani et al., NPJ Digital Medicine 1(1) 2018

DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

Discussion 1

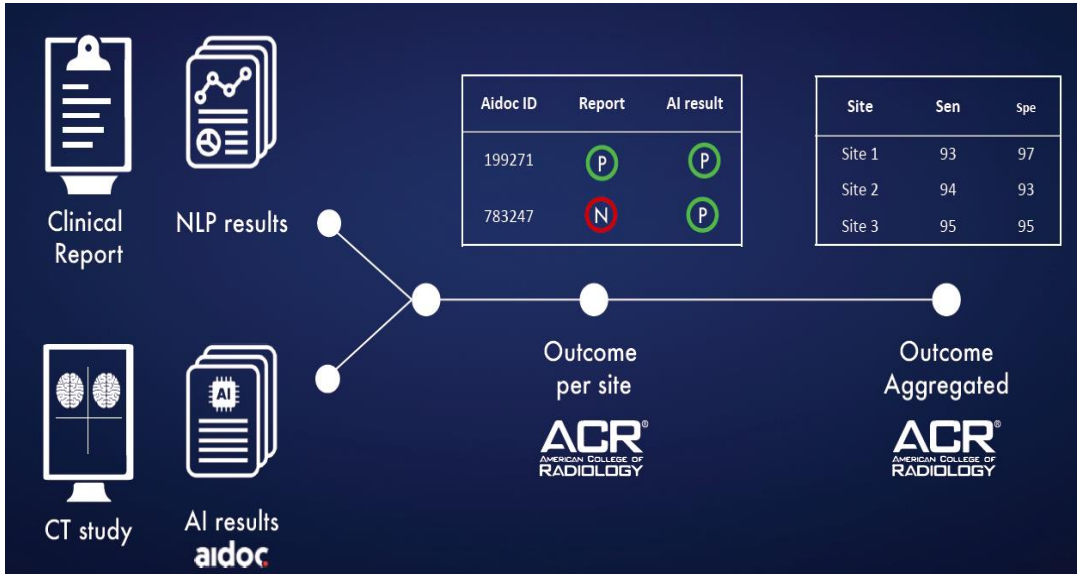
- We introduce a scientific framework, Artificial Intelligence Prospective Randomized Observer Blinding Evaluation (AI-PROBE) for quantitative clinical performance evaluation of radiology AI systems within prospective randomized clinical trials.

- Our evaluation workflow encompasses a study design and a corresponding radiology information technology infrastructure that randomly blinds radiologists with regards to the presence of positive reads as provided by AI-based image analysis systems.

Discussion 2

- AI-PROBE may be challenging for various technical, legal, and economic reasons.
- However, expected gains for many healthcare enterprise stakeholders, including patients, providers, hospitals, insurance companies, regulatory bodies (FDA, ACR), and others.
- Besides radiology turnaround times, other measurable quantities may be investigated, such as patient outcome or cost-effectiveness measures.
- Data collected during such trials can augment data repositories of regulartory bodies, such as the ACR Assess-AI initiative.

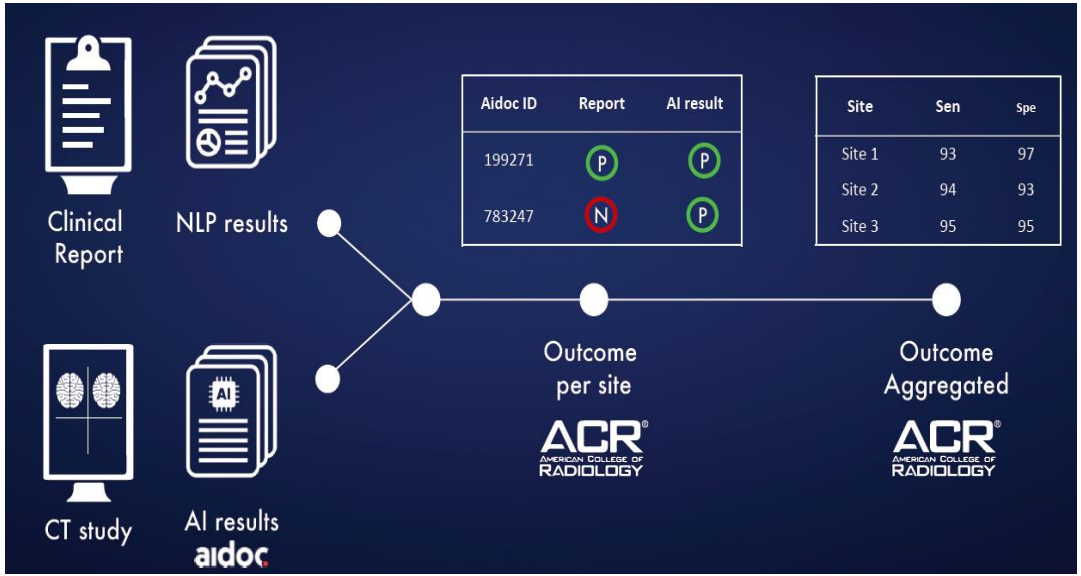
ACR Assess-AI: The Future



- Workflow architecture for data submission to the ACR Assess-AI repository for post-market surveillance of a vendor-specific AI solution (Aidoc, Tel Aviv, Israel).
- URMC was first institution in the US that went live submitting data to the ACR Assess-AI repository (January 2020)
- AI reading results are compared with NLP results of final radiology reports, serving as case-specific ground truth.
- Aggregated information can be used by ACR to monitor diagnostic accuracy of such AI solutions.

DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

ACR Assess-AI: The Future



We can augment the current ACR Assess-AI workflow by transmission of additional study-specific information, such as imaging meta-information, various time-stamps for measuring turn-around times, and potentially application-specific clinical outcome measures.

In the AI-PROBE framework, such extended data transmission will empower ACR to even perform virtual clinical trials comparing radiologists' performance with and without AI, thus creating an innovative infrastructure for substantially expanding ACR's capabilities for quantitative evaluation of AI solutions in clinical practice.

DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

Thank you for your attention!

DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY