# Retrospective comparison of AI algorithms

## Evaluation of 3 AI algorithms for intracranial hemorrhage

Kristoffer Järlevi, radiologist
Röntgenkliniken, VO Bilddiagnostik, Södersjukhuset

SÖS

SÖDERSJUKHUSET

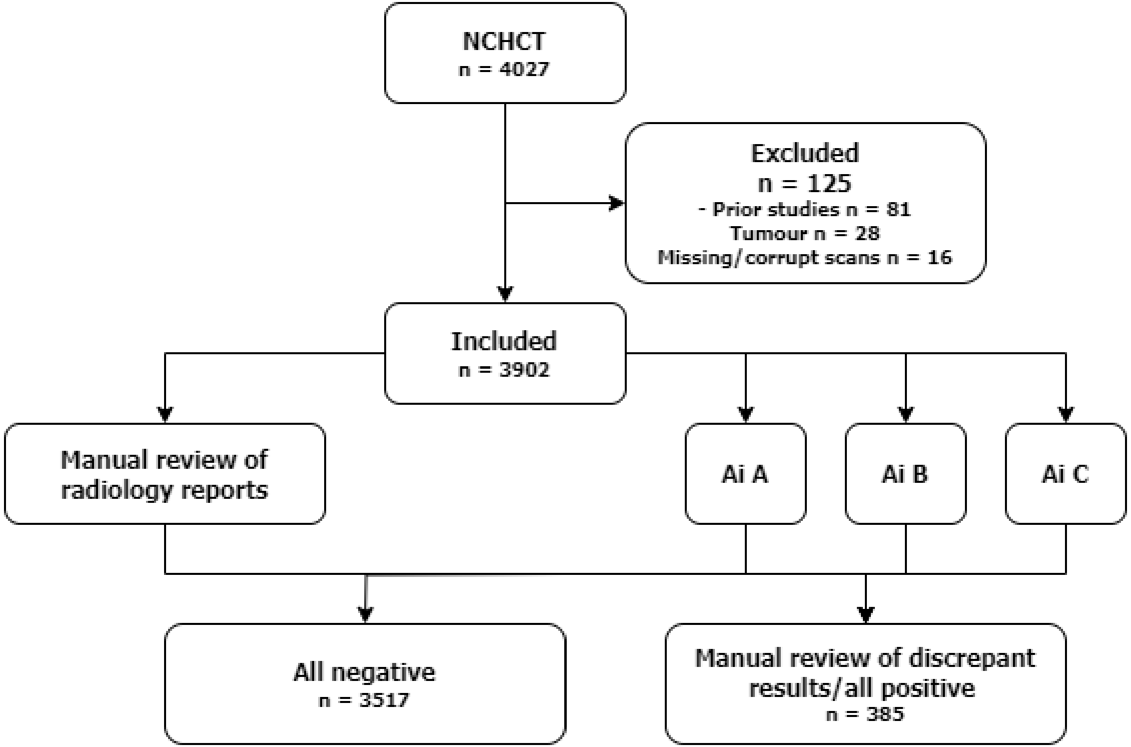
# Background

- Södersjukhuset in Stockholm, general hospital with high volumes (750 000 population).
  - Everything is double read
  - Operates 24/7 and many first interpretations are done by radiologists in training
- Increased interest in AI
- Where do we start?
- Are all algorithms equivalent? How do the algorithms perform in a 'first line' imaging setting?

# Method

- Retrospective study
- CT-head scans of adult patients during August-December 2022.
- Sensitivity, specificity, positive predictive value, and negative predictive value were calculated based on ground truth.
- 2 of the 3 vendors were anonymous

# Method – ground truth



NCHCT
n = 4027

Excluded
n = 125
- Prior studies n = 81
Tumour n = 28
Missing/corrupt scans n = 16

Included
n = 3902

Manual review of radiology reports

Ai A

Ai B

Ai C

All negative
n = 3517

Manual review of discrepant results/all positive
n = 385

# Results

- 3902 examinations were included
- 3726 negative for ICH
- 176 positive for ICH – prevalence of 4.5%
- 29 cases with ICH hade been overlooked by the first reader
- 8 cases were identified by AI that had been missed after double reading

# Results - demographics

| | Positive | | | Negative | | |
|---|---|---|---|---|---|---|
| Age | Male | Female | Unknown | Male | Female | Unknown |
| 18-29 | 4 | 3 | 0 | 138 | 134 | 1 |
| 30-39 | 4 | 0 | 0 | 163 | 151 | 5 |
| 40-49 | 5 | 2 | 0 | 152 | 155 | 3 |
| 50-59 | 13 | 9 | 0 | 193 | 213 | 2 |
| 60-69 | 16 | 6 | 0 | 277 | 240 | 7 |
| 70-79 | 29 | 15 | 0 | 420 | 405 | 3 |
| 80-89 | 27 | 22 | 0 | 344 | 420 | 3 |
| >90 | 8 | 13 | 0 | 94 | 203 | 0 |
| Total | 106 | 70 | 0 | 1781 | 1921 | 24 |

SÖS

SÖDERSJUKHUSET

# Results

| | |
|---|---|
| **Sensitivity** | Proportion of patients **with** a disease who test positive TP/(TP+FN) |

SÖDERSJUKHUSET

# Results

| | |
|---|---|
| **Sensitivity** | Proportion of patients **with** a disease who test positive TP/(TP+FN) |
| **Specificity** | Proportion of patients **without** a disease who test negative TN/(TN+FP) |

# Results

| | |
|---|---|
| **Sensitivity** | Proportion of patients **with** a disease who test positive TP/(TP+FN) |
| **Specificity** | Proportion of patients **without** a disease who test negative TN/(TN+FP) |
| **Positive Predictive Value** | Proportion of patients with a positive test who actually have the disease<br>TP/(TP+FP) |

# Results

| | |
|---|---|
| **Sensitivity** | Proportion of patients **with** a disease who test positive TP/(TP+FN) |
| **Specificity** | Proportion of patients **without** a disease who test negative TN/(TN+FP) |
| **Positive Predictive Value** | Proportion of patients with a positive test who actually have the disease TP/(TP+FP) |
| **Negative Predictive Value** | Proportion of patients with a negative test who actually dont have the disease TN/(TN+FN) |

# Results

| Interpreter | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|
| 1 Reader | 86,4% | 99,4% | 87,9% | 99,4% |
| 2 Reader | | | | |
| AI (A) | | | | |
| AI (B) | | | | |
| AI (C) | | | | |

# Results

| Interpreter | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|
| 1 Reader | 86,4% | 99,4% | 87,9% | 99,4% |
| 2 Reader | 95,5% | 99,5% | 90,8% | 99,8% |
| AI (A) | | | | |
| AI (B) | | | | |
| AI (C) | | | | |

# Results

| Interpreter | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|
| 1 Reader | 86,4% | 99,4% | 87,9% | 99,4% |
| 2 Reader | 95,5% | 99,5% | 90,8% | 99,8% |
| AI (A) | 61,4% | 97,2% | 50,7% | 98,2% |
| AI (B) | | | | |
| AI (C) | | | | |

# Results

| Interpreter | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|
| 1 Reader | 86,4% | 99,4% | 87,9% | 99,4% |
| 2 Reader | 95,5% | 99,5% | 90,8% | 99,8% |
| AI (A) | 61,4% | 97,2% | 50,7% | 98,2% |
| AI (B) | 63,6% | 97,5% | 54,4% | 98,3% |
| AI (C) | | | | |

# Results

| Interpreter | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|
| 1 Reader | 86,4% | 99,4% | 87,9% | 99,4% |
| 2 Reader | 95,5% | 99,5% | 90,8% | 99,8% |
| AI (A) | 61,4% | 97,2% | 50,7% | 98,2% |
| AI (B) | 63,6% | 97,5% | 54,4% | 98,3% |
| AI (C) | 90,3% | 99,0% | 80,3% | 99,5% |

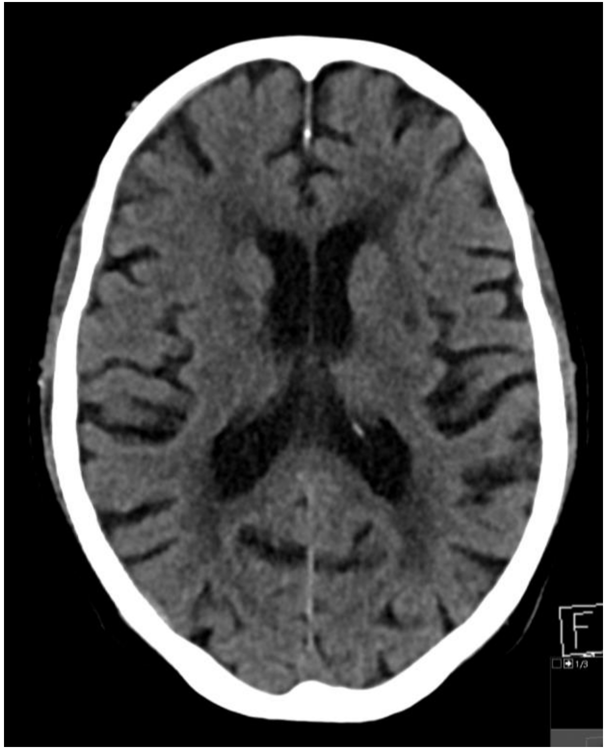
# Results – AI as support

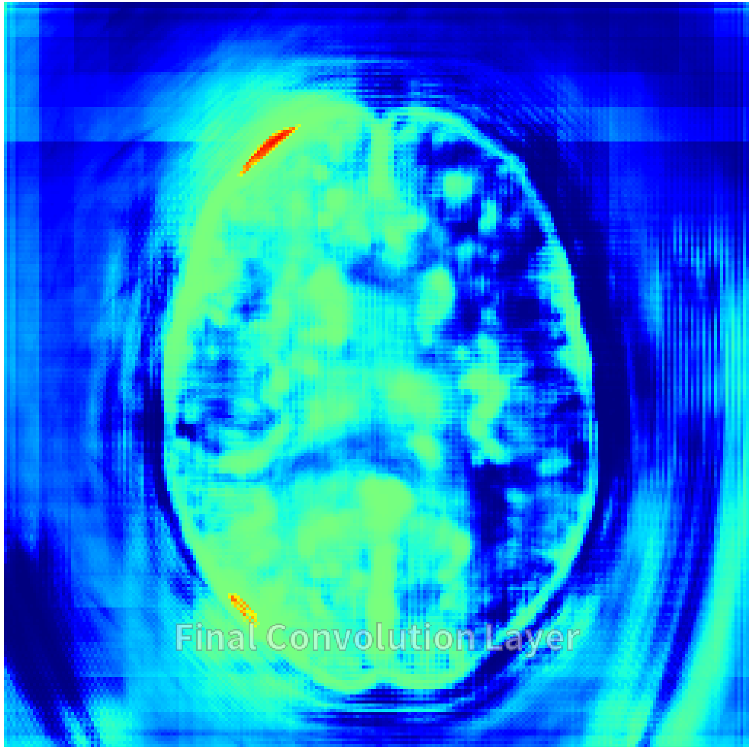
| Interpreter | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|
| 1 Reader | 86,4% | 99,4% | 87,9% | 99,4% |
| 2 Reader | 95,5% | 99,5% | 90,8% | 99,8% |
| 1 Reader + AI (C) | **96,0%** | 99,4% | 88,9% | 99,8% |
| 2 Reader + AI (C) | 98,9% | 99,5% | 91,1% | 99,9% |

# Results – test of superiority

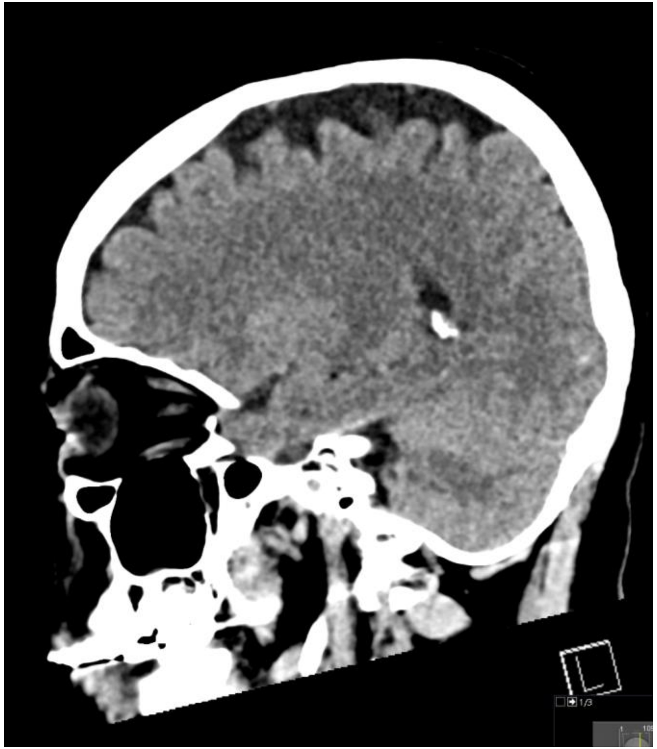
| Interpreter | Sensitivity | 1 Reader | 2 Reader | AI (C) | 1 Reader + AI (C) |
|---|---|---|---|---|---|
| **1 Reader** | 86,4% |  | 1 | 0,94 | 1 |
| **2 Reader** | 95,5% | <0,001 |  | 0,039 | 0,73 |
| **AI (C)** | 90,3% | 0,12 | 1 |  | 1 |
| **1 Reader + AI (C)** | 96,0% | **<0,001** | 0,50 | 0,001 |  |

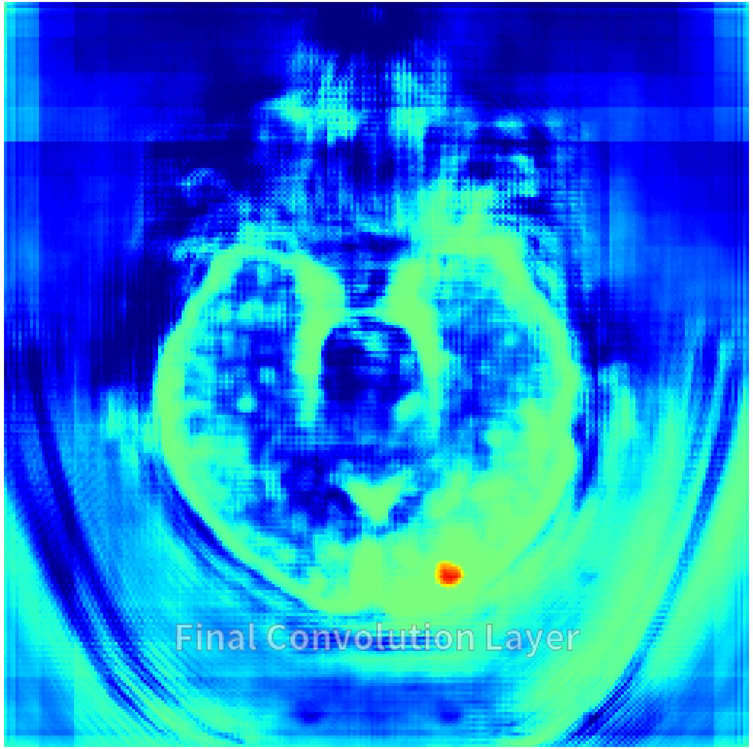
# Results – False negative radiologists 1



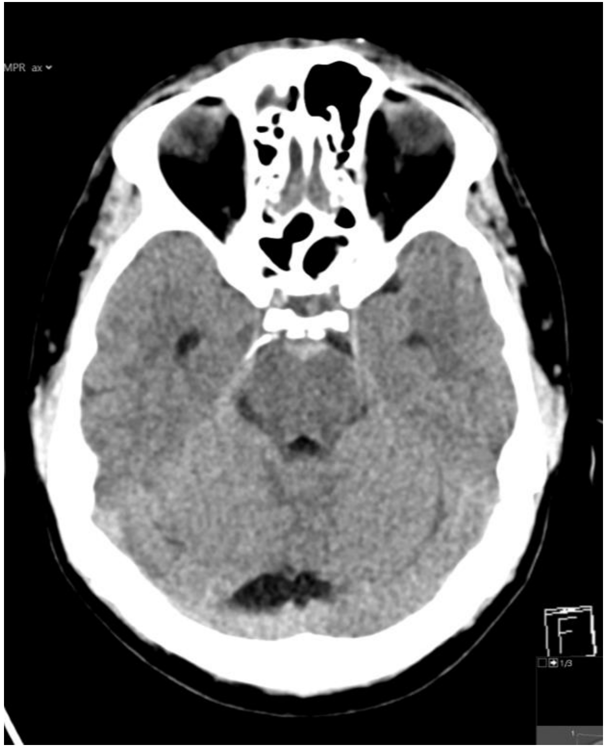


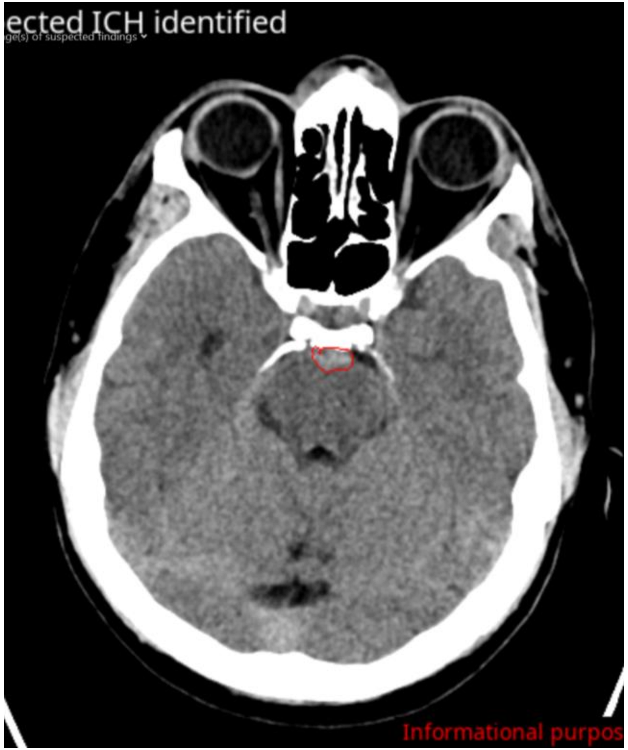
# Results – False negative radiologists 2





# Results – False negative radiologists 3





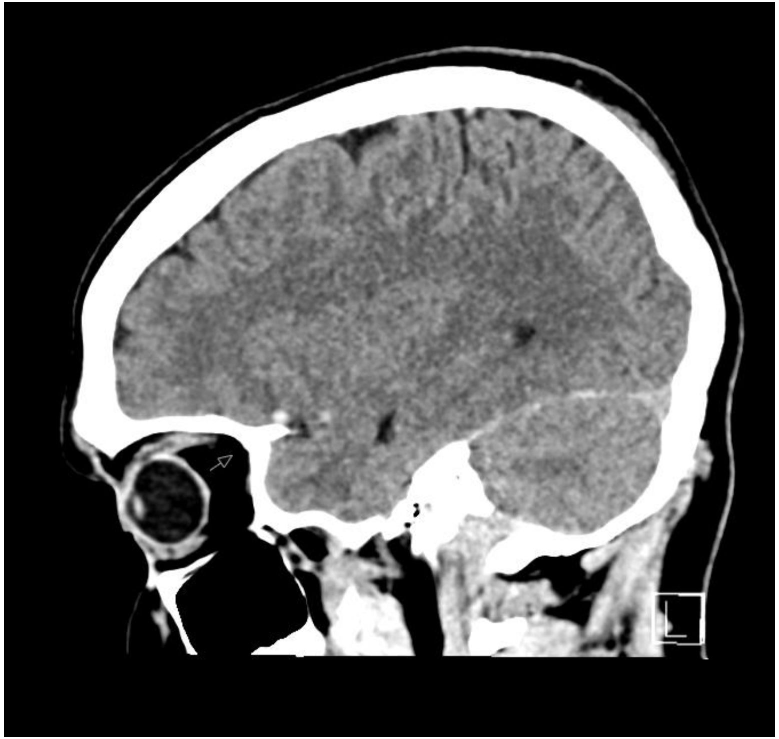
SÖS
SÖDERSJUKHUSET

# Results – False negative AI (C)

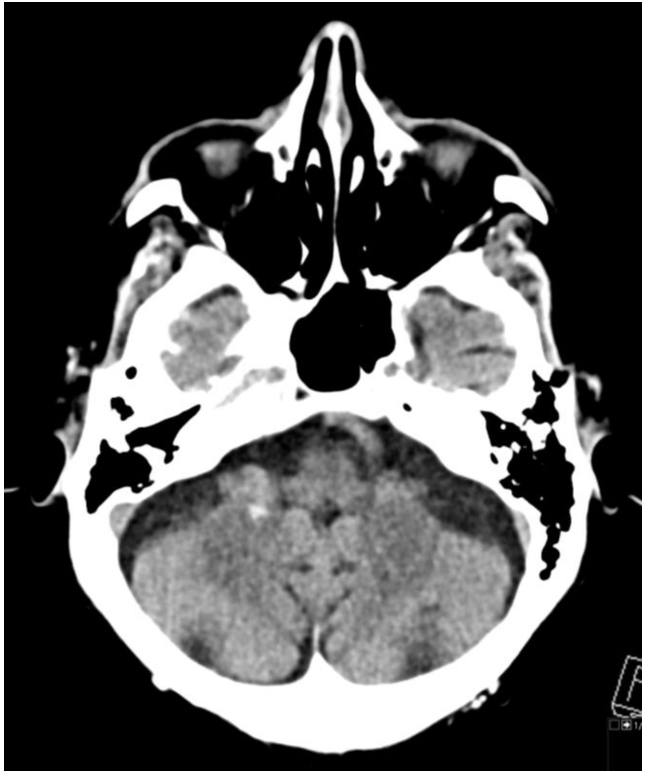
- 17

  - 6 chronic subdural hematoma with acute components
  - 5 subarachnoid hemorraghes
  - 3 contusions
  - 1 parenchymal hemorraghe
  - 2 scans with several hemorraghes

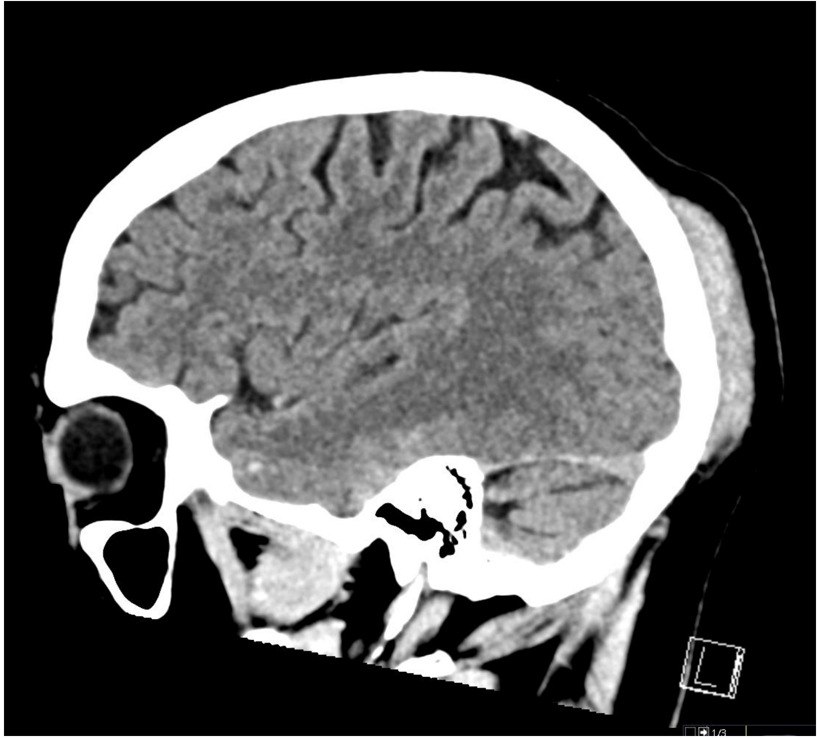
# Results – False negative AI (C)



# Results – False negative AI (C)



# Results – False negative AI (C)



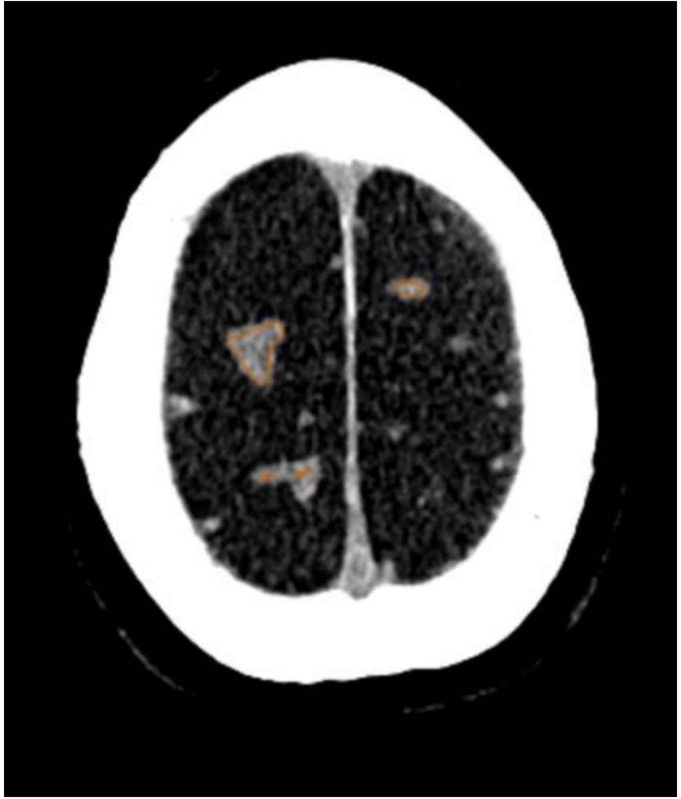
# Results – type of ICH

| Type | Number of ICH | AI (A) % | AI (B) % | AI (C) % |
|---|---|---|---|---|
| Subarachnoidal | 96 | 55% | 57% | 82% |
| Subdural | 88 | 57% | 58% | 75% |
| Contusion | 28 | 46% | 57% | 64% |
| Parenchymal | 43 | 70% | 70% | 81% |
| Epidural | 1 | 0% | 0% | 0% |
| **Total** | **256** | **57%** | **59%** | **77%** |

# Results – AI false positive

| False Positive | AI (A) | AI (B) | AI (C) | Total | Total % |
|---|---|---|---|---|---|
| Falx/sinus | **85** | 7 | 7 | 99 | 40% |
| Meningeoma | 7 | 10 | **14** | 31 | 13% |
| Parenchyma | 4 | **23** | 2 | 29 | 12% |
| Op-material/dura | 2 | 6 | 4 | 12 | 5% |
| Artifact | 6 | 11 | 0 | 17 | 7% |
| Vessels | 2 | 2 | 1 | 5 | 2% |
| Calcifications | 1 | 6 | 6 | 13 | 5% |
| Unidentified | 0 | 17 | 0 | 17 | 7% |
| Other | 3 | 16 | 5 | 24 | 10% |
|  | 110 | 98 | 39 | 247 |  |

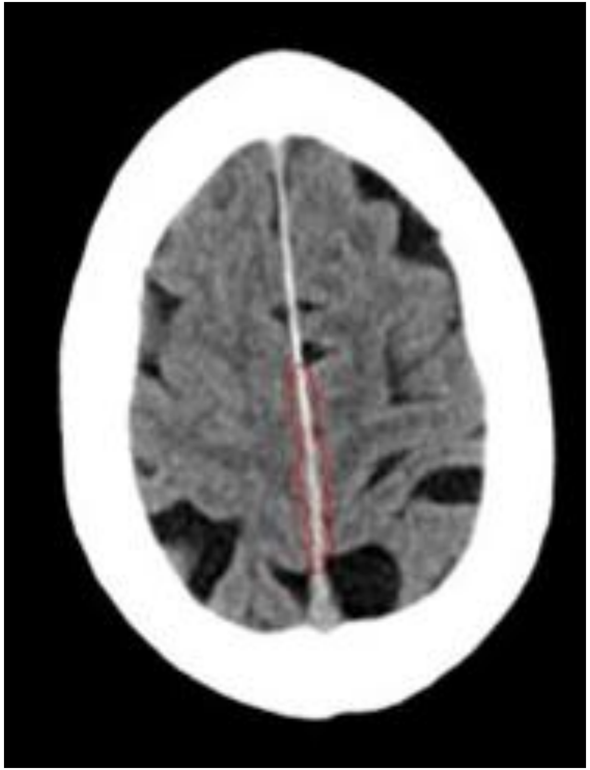
# Results – false positive 1



SÖS

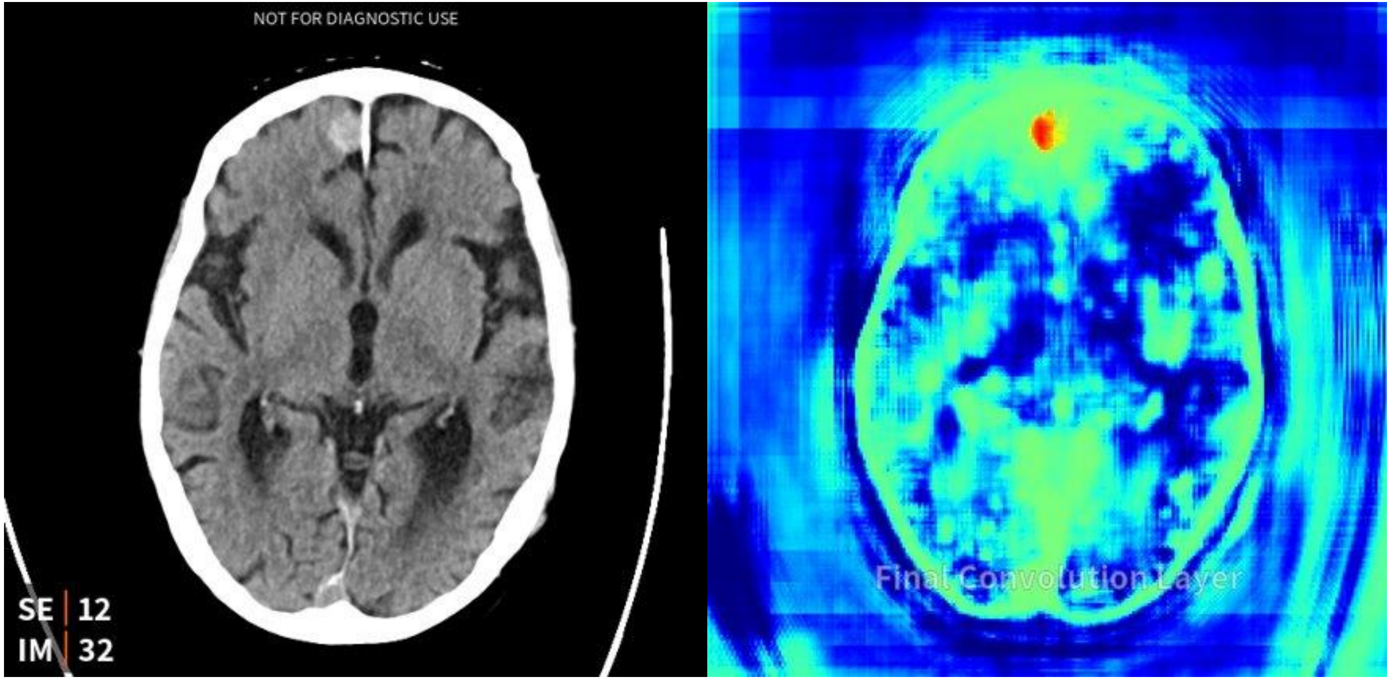
SÖDERSJUKHUSET

# Results – false positive 2



SÖDERSJUKHUSET

# Results – false positive 3



# Results – time of day

| Interpreter | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|
| 1 Reader | 86,5% | 99,4 % | 88,0% | 99,4% |
| 2 Reader | 95,5% | 99,6% | 90,9% | 99,8% |
| On-call hours (15:30-07) | 87,9% | 99,4% | 87,2% | 99,4% |
| Single reader (21-07) | 89,4% | 99,5% | 87,5% | 99,6% |
| Regular hours (07-15:30) | 82,6% | 99,6% | 90,5% | 99,2% |

# Results – time of day

| Interpreter | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|
| 1 Reader | 86,5% | 99,4 % | 88,0% | 99,4% |
| 2 Reader | 95,5% | 99,6% | 90,9% | 99,8% |
| On-call hours (15:30-07) | 87,9% | 99,4% | 87,2% | 99,4% |
| Single reader (21-07) | **89,4%** | 99,5% | 87,5% | 99,6% |
| Regular hours (07-15:30) | **82,6%** | 99,6% | 90,5% | 99,2% |

# Results – time of day

| Interpreter | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|
| 1 Reader | 86,5% | 99,4 % | 88,0% | 99,4% |
| 2 Reader | 95,5% | 99,6% | 9[illegible],9% | 9[illegible],[illegible]% |
| On-call hours (15:30-07) | 87,9% | 99,4% | 87,2% | 99,4% |
| Single reader (21-07) | **89,4%** | 99,5% | 87,5% | 99,6% |
| Regular hours (07-15:30) | **82,6%** | 99,6% | 90,5% | 99,2% |

**Cut back on radiologists?**

# Discussion

- ICH potential serious consequences
  - Sensitivity and NPV are the most important
  - AI C had higher sensitivity and same NPV as 1 Reader
  - AI C found 24 TP that 1 Reader overlooked.
    - Typically small ICH.
- AI C had lower PPV
  - Often easily dismissible by a human radiologist

# Discussion

- Combining AI with 1 Reader
  - Sensitivity increased from 86 to 96%.
    - AI identififying more true positives
  - PPV improved from 80% to 88%.
    - Radiologist generating fewer false positives
- AI + 1 Reader vs double reading
  - Were not able to show superior performance
  - Provide real-time feedback to 1 Reader → alter clinical management
  - Can not replace radiologist
  - Prospective study

# Discussion

- The only (?) study comparing different algorithms on same dataset

# Discussion

- The only (?) study comparing different algorithms on same dataset
- Large variation between the algorithms

# Discussion

- The only (?) study comparing different algorithms on same dataset
- Large variation between the algorithms
- Only one algorithm performed on par with the 1 Reader

# Discussion

- The only (?) study comparing different algorithms on same dataset
- Large variation between the algorithms
- Only one algorithm performed on par with the 1 Reader
- CE and FDA emphasize safety over efficacy

# Discussion

- The only (?) study comparing different algorithms on same dataset
- Large variation between the algorithms
- Only one algorithm performed on par with the 1 Reader
- CE and FDA emphasize safety over efficacy
- Need of doing one’s own validation

# Tack!

SÖS

SÖDERSJUKHUSET