## Appendix – Quantitative results

In the primary analysis, AI use consistently increased decision-making and total reading times regardless of which worklist had the AI output. When the x-ray classification tool was made available for Worklist X, users took about 3 seconds longer (20% increase) to start dictating ($p = 0.02$) and 23 seconds longer (45% increase) to complete dictations ($p = 0.18$). When the x-ray classification tool was made available for Worklist Y, users took about 7 seconds longer (84% increase) to start dictating ($p = 0.09$) and 5 seconds longer (14% increase) to complete dictations ($p = 0.19$).

**Supplementary Table 1. Worklist case selection for Round 2**

| Case # | Worklist (X or Y) | Actual diagnosis (verified by the principal investigator) | AI binary label | Confidence level | Comments |
|---|---|---|---|---|---|
| 1 | X | Normal | Normal | 0 | |
| 2 | Y | Normal | Normal | 18 | |
| 3 | X | Normal | Normal | 0 | |
| 4 | Y | Normal | Normal | 0 | |
| 5 | X | Normal | Normal | 0 | |
| 6 | Y | Normal | Normal | 5 | |
| 7 | X | Normal | Normal | 0 | |
| 8 | Y | Normal | Normal | 15 | |
| 9 | X | Normal | Normal | 11 | |
| 10 | Y | Normal | Normal | 0 | |
| 11 | X | Normal | Normal | 5 | |
| 12 | Y | Normal | Normal | 0 | |
| 13 | X | Normal | Normal | 0 | |
| 14 | Y | Normal | Normal | 9 | |
| 15 | X | Normal | Normal | 0 | |
| 16 | Y | Normal | Normal | 0 | |
| 17 | X | Normal | Abnormal | 26 | False positive |
| 18 | Y | Normal | Abnormal | 51 | False positive |
| 19 | X | Abnormal - pneumothorax & pleural effusion | Abnormal | 95 | Used to assess time to reach a critical case |
| 20 | Y | Abnormal - pneumothorax & pleural effusion | Abnormal | 95 | Used to assess time to reach a critical case |
| 21 | X | Abnormal - lobar pneumonia | Abnormal | 81 | |
| 22 | Y | Abnormal - lobar pneumonia | Abnormal | 95 | |
| 23 | X | Abnormal - atelectasis | Abnormal | 74 | |
| 24 | Y | Abnormal - atelectasis | Abnormal | 11 | |
| 25 | X | Abnormal - diffuse sclerotic bone lesions | Abnormal | 89 | |
| 26 | Y | Abnormal - diffuse sclerotic bone lesions | Abnormal | 76 | |
| 27 | X | Abnormal - cavitary lesion | Abnormal | 95 | |
| 28 | Y | Abnormal - cavitary lesion | Abnormal | 95 | |
| 29 | X | Abnormal - pleural effusion, pneumoperitoneum, & atelectasis | Abnormal | 93 | |
| 30 | Y | Abnormal - pleural effusion, pneumoperitoneum, & airspace opacification | Abnormal | 86 | |
| 31 | X | Abnormal - nodules | Abnormal | 85 | |
| 32 | Y | Abnormal - nodules | Abnormal | 52 | |
| 33 | X | Abnormal - hiatal hernia | Abnormal* | 83 | False negative |
| 34 | Y | Abnormal - hiatal hernia | Normal* | 7 | False negative |

*Both Worklist X and Worklist Y contained a case of hiatal hernia. Although the x-ray classification tool labeled the hiatal hernia case as abnormal in Worklist X and normal in Worklist Y, both cases were considered to be false negatives for our study, as the heatmap did not mark the area with the hiatal hernia.

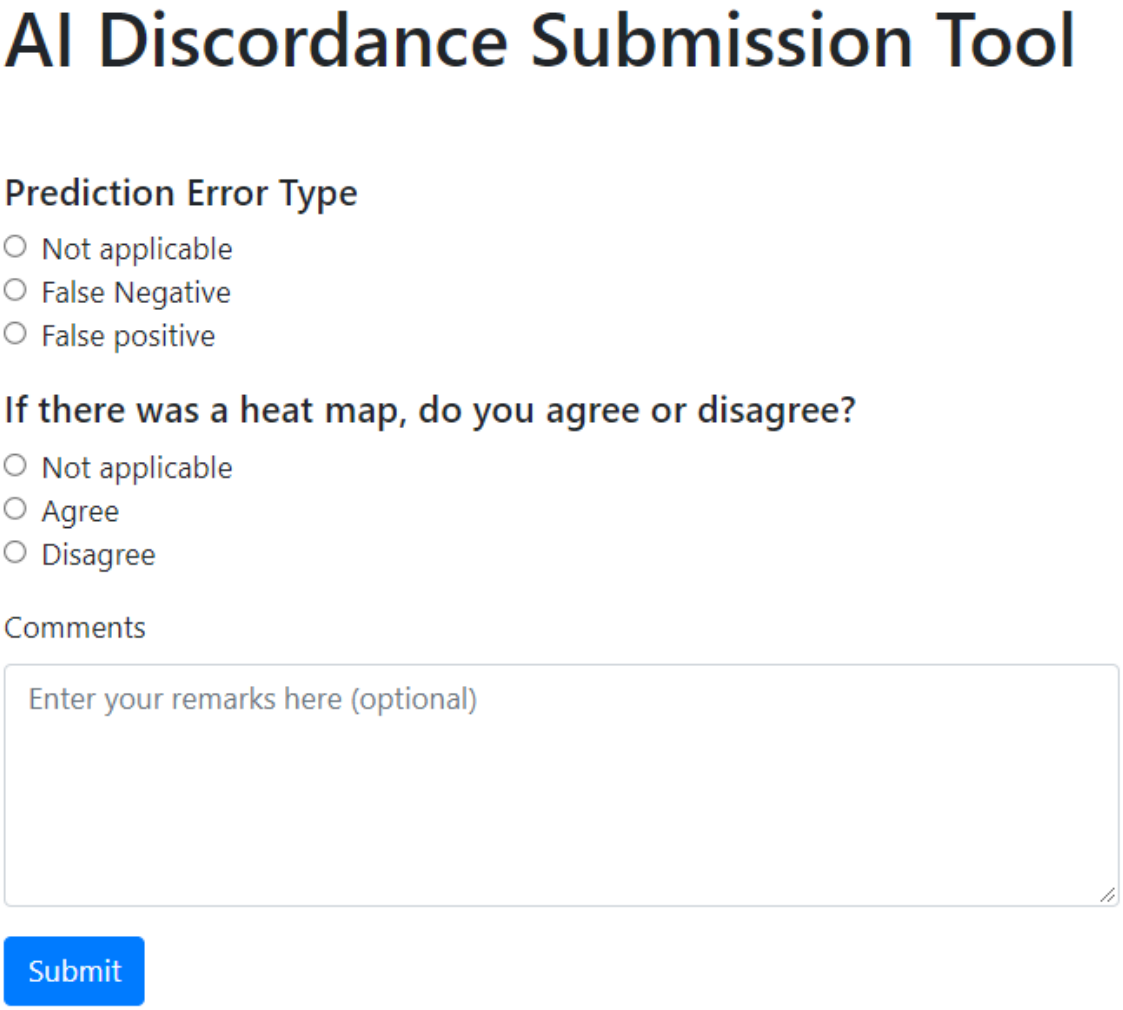
**Supplementary Table 2. Quantitative results from round 2.**

| Participant | Worklist with AI (X or Y) | Was triage used? | Average time (seconds) to start dictating a case with AI (95% CI) | Average time (seconds) to start dictating a case without AI (95% CI) | p value | Average time (seconds) to complete a case with AI (95% CI) | Average time (seconds) to complete a case without AI (95% CI) | p value |
|---|---|---|---|---|---|---|---|---|
| P1 | X | no | 17.56 (12.11, 23.01) | 13.65 ( 7.67, 19.63) | 0.33 | 44.75 ( 32.74, 56.76) | 35.41 (20.73, 50.09) | **0.05** |
| P2 | X | yes | 22.00 (17.29, 26.71) | 18.65 (14.03, 23.26) | 0.31 | 53.71 ( 36.56, 70.85) | 45.59 (34.42, 56.75) | 0.33 |
| P3 | X | yes | 14.59 (11.38, 17.79) | 10.76 ( 8.79, 12.74) | **0.05** | 35.76 ( 26.08, 45.45) | 28.24 (22.30, 34.17) | 0.09 |
| | X | yes | | | | 191.8 | | |
| P4 | | | 28.89 (13.02, 37.86) | 29.00 (19.14, 38.86) | 0.21 | 9 (141.82, 241.95) | 113.85 (81.89, 145.81) | 0.12 |
| P5 | Y | yes | 18.12 ( 9.71, 26.53) | 10.29 ( 4.35, 16.24) | 0.16 | 47.71 ( 29.54, 65.87) | 38.59 (23.75, 53.42) | **0.04** |
| P6 | Y | yes | 28.06 ( 9.87, 46.25) | 8.35 ( 3.89, 12.24) | **0.03** | 47.53 ( 27.60, 67.46) | 34.41 (20.60, 48.22) | **0.03** |
| P7 | Y | no | 12.35 ( 5.52, 19.18) | 7.88 ( 5.90, 9.86) | 0.19 | 22.18 ( 14.30, 30.05) | 19.18 (13.17, 25.18) | 0.49 |
| P8 | Y | no | 14.00 (10.72, 17.28) | 10.29 ( 7.80, 12.79) | **0.01** | 56.88 ( 38.35, 75.42) | 54.12 (36.26, 71.98) | 0.65 |
| P9 | Y | yes | 7.65 ( 5.66, 9.63) | 6.88 ( 3.72, 10.05) | 0.52 | 18.94 ( 13.85, 24.03) | 23.35 (14.52, 32.19) | 0.14 |
| P10 | X | no | 12.06 ( 8.84, 15.28) | 7.47 ( 4.19, 10.75) | **0.05** | 40.82 ( 26.10, 55.55) | 30.18 (19.09, 41.26) | **0.04** |
| Overall | | | 17.53 (12.55, 22.50) | 12.32 ( 7.44, 17.21) | **0.02** | 56.02 ( 20.73, 91.30) | 42.29 (22.87, 61.71) | 0.09 |

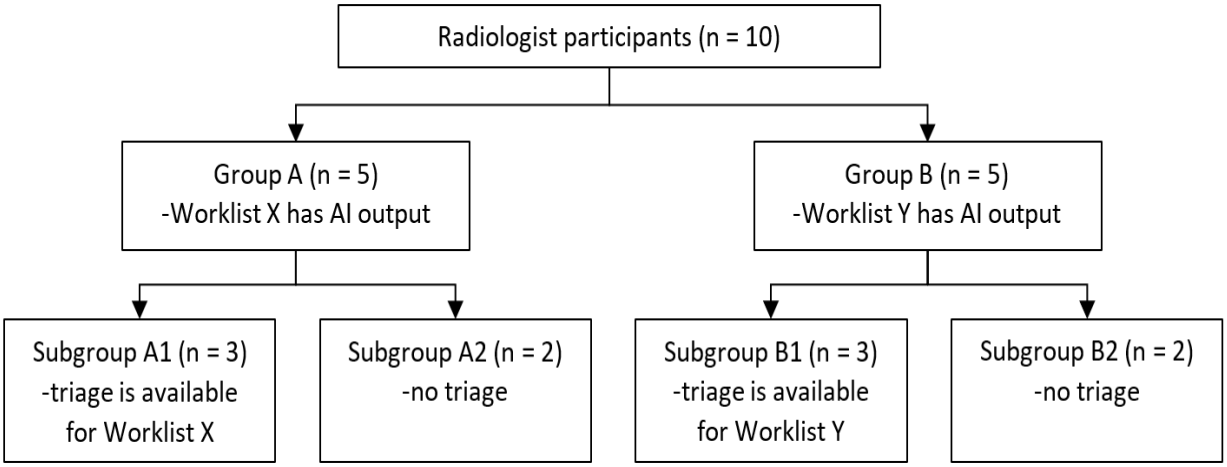
**Supplementary Table 3. Quantitative results from round 2 (sensitivity analysis to account for a learning curve*)**

| Participant | Worklist with AI (X or Y) | Was triage used? | Average time (seconds) to start dictating a case with AI (95% CI) | Average time (seconds) to start dictating a case without AI (95% CI) | p value | Average time (seconds) to complete a case with AI (95% CI) | Average time (seconds) to complete a case without AI (95% CI) | p value |
|---|---|---|---|---|---|---|---|---|
| P1 | X | no | 16.80 (11.20, 22.40) | 14.27 ( 7.49, 21.04) | 0.49 | 43.80 ( 31.07, 56.53) | 30.20 ( 16.73, 43.67) | 0.07 |
| P2 | X | yes | 21.63 (16.66, 26.59) | 19.50 (14.96, 24.04) | 0.51 | 48.25 ( 34.71, 61.79) | 43.44 ( 32.53, 54.35) | 0.55 |
| P3 | X | yes | 13.56 (11.04, 16.08) | 10.94 ( 8.86, 13.02) | 0.10 | 33.38 ( 24.54, 42.21) | 28.06 ( 21.72, 34.40) | 0.17 |
| P4 | X | yes | 30.57 (30.57, 42.22) | 25.00 (11.78, 38.22) | 0.35 | 184.29 (133.39, 235.18) | 151.29 (124.71, 177.86) | 0.20 |
| P5 | Y | yes | 15.38 ( 8.87, 21.88) | 10.50 ( 4.15, 16.85) | 0.31 | 44.44 ( 26.46, 62.41) | 38.13 ( 22.28, 53.97) | 0.06 |
| P6 | Y | yes | 28.50 ( 9.05, 47.95) | 7.75 ( 3.93, 11.68) | **0.03** | 47.19 ( 25.86, 68.51) | 31.81 ( 18.26, 45.37) | **0.01** |
| P7 | Y | no | 12.47 ( 5.17, 19.77) | 8.00 ( 5.90, 10.10) | 0.21 | 22.03 ( 13.62, 30.45) | 19.00 ( 12.59, 25.41) | 0.50 |
| P8 | Y | no | 13.75 (10.29, 17.21) | 10.56 ( 7.97, 13.16) | **0.02** | 52.81 ( 35.25, 70.37) | 53.75 ( 34.65, 72.85) | 0.85 |
| P9 | Y | yes | 7.31 ( 5.33, 9.29) | 7.00 ( 3.62, 10.38) | 0.79 | 18.63 ( 13.22, 24.03) | 22.69 ( 13.35, 32.02) | 0.20 |
| P10 | X | no | 11.88 ( 8.46, 15.29) | 7.69 ( 4.21, 11.17) | 0.08 | 39.81 ( 24.22, 55.41) | 30.56 ( 18.73, 42.39) | 0.07 |
| Overall | | | 17.18 (11.84, 22.53) | 12.12 ( 7.90, 16.34) | **0.02** | 53.46 (19.62, 87.31) | 44.89 ( 17.20, 72.58) | **0.03** |

*A sensitivity analysis was performed by removing the first case from the worklist where the AI results were available to account for a learning curve




**Supplementary Figure 1. Discordance reporting page.** The discordance reporting page includes optional radio buttons for the user to categorize the perceived error in the AI output. There is also an optional free-text field for the users to explain their rationale.




**Supplementary Figure 2. Flow diagram of Round 2 participants.** A total of 10 radiologists were recruited. Users were randomly assigned to one of two groups (Group A or B), and both groups reviewed the two worklists (X and Y). To compare the average time to begin dictating and to completely read a single x-ray (with vs without AI assistance), the x-ray classification tool was available for one worklist (Worklist X for Group A and Worklist Y for B users). To assess whether the triage function would impact the time to reach a critical finding for a given worklist, users were further stratified into subgroups such that Subgroups A1 and B1 had the triage function available for their corresponding worklist with the AI output, and Subgroups A2 and B2 did not have the triage function available at any point in the testing.

1. Please tell us how you felt about using the tool in general.
2. Do you think this tool could be useful in a real-world workflow?
3. Would you personally use this tool in your clinical work?
4. What are your thoughts on the different features of this tool?
5. Were there aspects of this tool that you found frustrating to use?
6. Out of the two tool layouts (I.e., with and without the heat map present) which did you prefer? (Only asked in Round 1)
7. How has this tool impacted the way you interpreted the images?
8. How has this tool affected your confidence in your interpretation of the images?
9. What parts of the tool, if any, would you change to better fit your needs?
10. On a scale of 1 to 5, (1 being the worst and 5 being the best), how would you rate the performance of this tool overall?
11. Do you have anything else that you would like to mention about your experience?

**Supplementary Figure 3. Post-testing question guide.** These questions were used as a starting point for the post-testing interview. When radiologists introduced new topics or provided vague responses, probing questions were used to elicit further details.