ARTERYS

Bridging the gap

How Arterys Lung AI improves inter-radiologist agreement

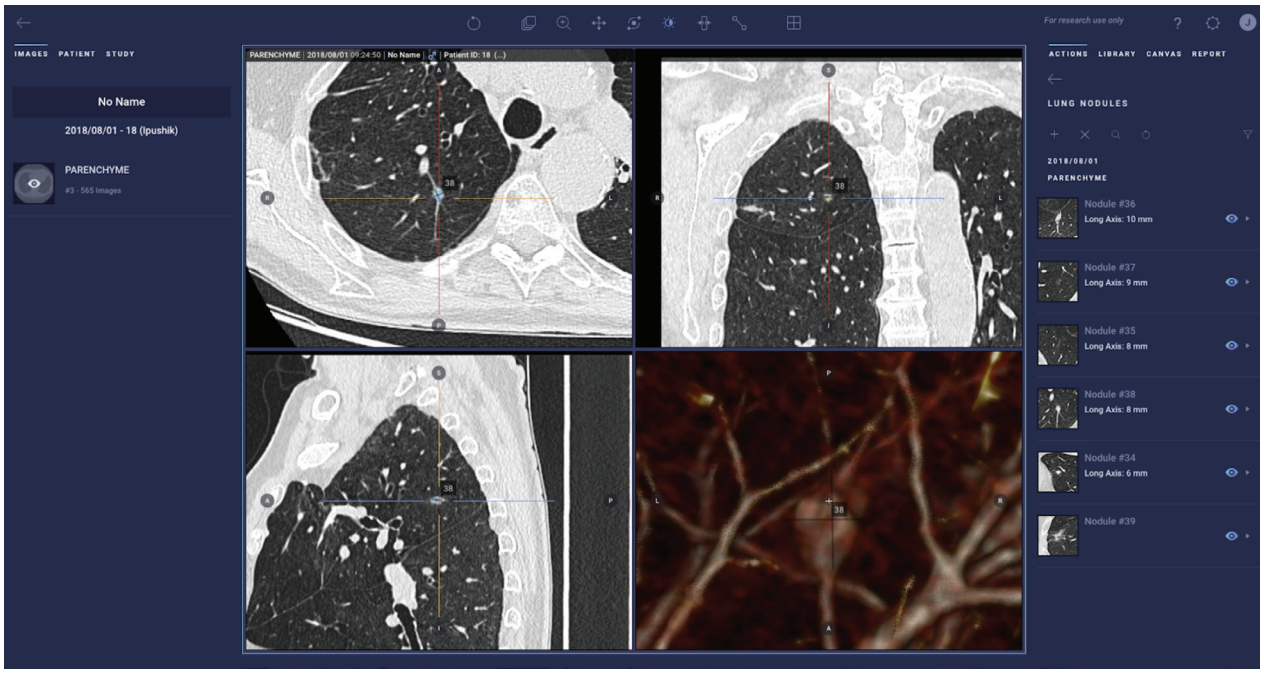
Introduction

A multi-center, multi-reader retrospective study was designed to investigate if reads augmented by Arterys Lung AI1 detection improve inter-radiologist agreement.

Materials & Methods

A set of thirty thoracic CTs were randomly selected from two European hospitals[2] [3]. Three selection categories were defined based on clinical reports: screening examination with nodules, screening examinations without nodules, and other pathology/incidentals with unknown nodules. Within each category, 5 studies were randomly selected from each institution.

Three radiologists from three separate institutions read all studies twice; with and without Arterys Lung AI detection. The reading order was randomized and a cool-off period of 10 days was set between the reads.

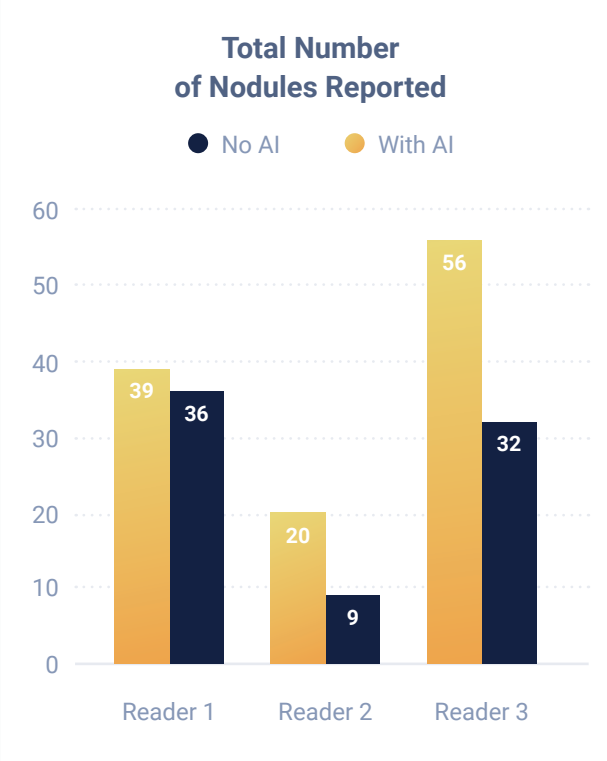


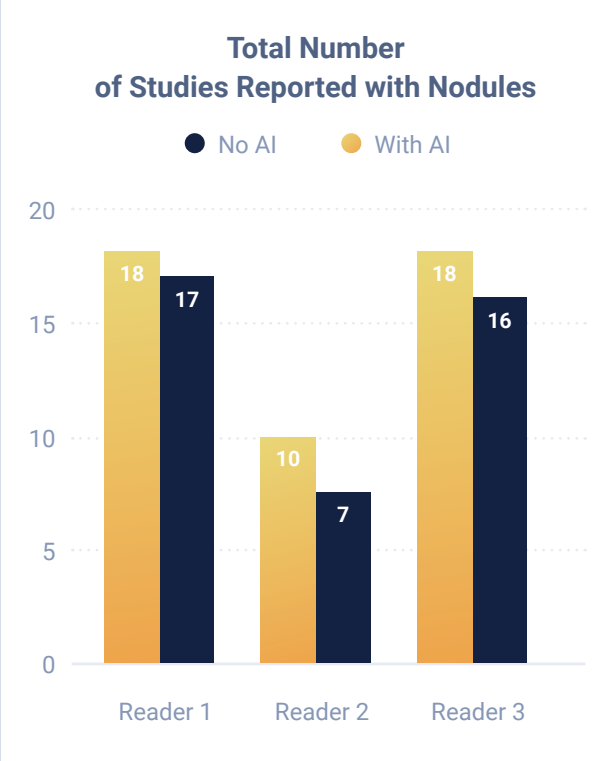
For the reads without detection, the radiologist was instructed to add nodules, using a nodule auto-segmentation tool, requiring a single click per nodule from the user. For the reads using detection, detected nodules were reviewed by the radiologist and were either accepted or rejected (by deletion). Any missed nodule was added using the nodule auto-segmentation tool. Each radiologist was instructed to keep a spreadsheet to track the number of added nodules and deleted nodules (for detection only).

Figure 1.0: The Arterys Lung AI application displaying four views of a nodule; axial, coronal, sagittal and direct volume rendering. Alongside a list of detected nodules shown on the right.• Arterys maintains the Edge server

Results

The use of AI reduced the variation in the total number of reported locations of concern (clinically actionable nodules). It increased the total number of reported nodules as well as studies reported to have at least one nodule (Graph 1.0 and 2.0). We note differences of 4x among radiologists when not using AI compared to 2.8x when using AI. For the number of studies reported containing nodules, the differences among radiologists were 2.4x when not using AI compared to 1.8x when using AI.

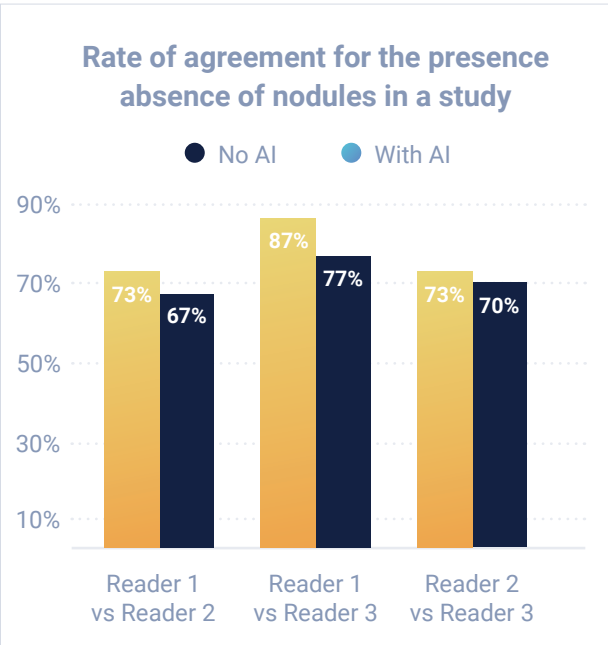




Graph 1.0 and 2.0: Total number of reported nodules and reported studies containing nodules per reader

The percentage agreement was evaluated between all radiologist pairs. Agreement was defined as true if both radiologists reported at least one nodule in the study or if both radiologists agreed there were no nodules present.

Agreement was as low as 67% when not using AI and as high as 87% when using AI (Graph 3.0). In all cases, using AI improved the agreement rate between individual radiologists for the presence or absence of nodules in a study.



Graph 3.0: Total number of reported nodules and reported studies containing nodules per reader

A further measure of agreement called the multi-rater kappa (Fleiss' kappa[4]) was used. Multi-rater kappa allows to take into account the possibility that the amount of agreement could be expected to occur through chance.

A useful guideline to help interpret kappa is that a value of 0.0 will represent no agreement and a value of 1.0 will represent perfect agreement.

Multi-rater kappa ranged from 0.42 (95% confidence interval: 0.31, 0.52) when not using AI to 0.56 (95% confidence interval: 0.45, 0.66) when using AI, see Graph 4.0. Using AI improved radiologist agreement from moderate to substantial[5].

Multi-Rater Kappa

	With AI	No AI
k	0.56	0.42
SE (k)	0.11	0.11
z-score	5.27	3.94
p-value	< 0.0001	< 0.0001

Graph 4.0: Multi-Rater agreement for the presence or absence of nodules in a study.

Conclusion

The study demonstrates how augmenting the radiologists with Arterys Lung AI[1] detection can reduce reporting variability. Using AI, radiologists missed less studies containing nodules and also included more nodules for further monitoring and investigation.

1. Arterys Lung AI v18.09
2. Centre Hospitalier de Valenciennes, Avenue Désandrouin CS 50479, 59322 Valenciennes, FRANCE
3. Groupe hospitalier Paris Saint-Joseph, 185, rue Raymond Losserand - 75014 Paris, FRANCE
4. Fleiss, J. L. (1971) "Measuring nominal scale agreement among many raters." Psychological Bulletin, Vol. 76, No. 5 pp. 378–382
5. Landis, J.R.; Koch, G.G. (1977). "The measurement of observer agreement for categorical data". Biometrics. 33 (1): 159–174

ARTERYS

info@arterys.com
51 Federal St. Suite 305, San Francisco, CA 94107
www.arterys.com