



OPEN

DATA DESCRIPTOR

A Dataset for Understanding Radiologist-Artificial Intelligence Collaboration

Alex Moehring¹✉, Manasi Kutwal², Ray Huang², Oishi Banerjee³, Adam Jacobi⁴, Corey Eber⁴, Dexter Mendoza⁴, Mike Chung⁴, Etan Dayan⁴, Yogesh Gupta⁵, Tan D. T. Bui⁶, Steven Q. H. Truong⁶, Anuj Pareek^{7,8}, Curtis P. Langlotz⁹, Matthew P. Lungren^{10,11,12}, Nikhil Agarwal^{13,14}, Pranav Rajpurkar^{3,14} & Tobias Salz^{13,14}

This dataset, Collab-CXR, provides a unique resource to study human-AI collaboration in chest X-ray interpretation. We present experimentally generated data from 227 professional radiologists who assessed 324 historical cases under varying information conditions: with and without AI assistance, and with and without clinical history. Using a custom-designed interface, we collected probabilistic assessments for 104 thoracic pathologies using a comprehensive hierarchical reporting structure. This dataset is the largest known comparison of human-AI collaborative performance to either AI or humans alone in radiology, offering assessments across an extensive range of pathologies with rich metadata on radiologist characteristics and decision-making processes. Multiple experimental designs enable both within-subject and between-subject analyses. Researchers can leverage this dataset to investigate how radiologists incorporate AI assistance, factors influencing collaborative effectiveness, and impacts on diagnostic accuracy, speed, and confidence across different cases and pathologies. By enabling rigorous study of human-AI integration in clinical workflows, this dataset can inform AI tool development, implementation strategies, and ultimately improve patient care through optimized collaboration in medical imaging.

Background & Summary

AI has matched or exceeded human ability in many domains (e.g. complex games like Go and Chess, speech recognition, art imitation, etc.) However, it still cannot execute the full range of tasks that humans can¹. There is great interest in understanding the drivers of whether AI tools outperform humans, whether human-AI collaboration exceeds humans alone, and in designing collaborative setups that blend the complementary skill sets of humans and AI^{2–4}. The answers to these questions may also depend on factors such as human skill and AI quality. In the context of radiology, AI currently excels when interpreting images in isolation, yet humans may be able to integrate additional information from a patient's history that is not yet incorporated into the AI algorithms. To better understand these questions for radiology applications, there is a need for experimental data that measures diagnostic performance under varying collaborative environments.

The Collab-CXR dataset⁵ is generated from an experiment first analyzed and reported in⁶, which collects chest X-ray diagnostic assessments for retrospective patient cases of 227 radiologists (also referred to as participants or experts) for 104 pathologies using a remote interface designed for this study. We collect data under different information conditions and experimental designs, making it possible to determine how AI assistance and

¹Purdue University, Daniels School of Business, West Lafayette, IN, 47907, US. ²MIT Economics, Blueprint Labs, Cambridge, MA, 02142, US. ³Harvard Medical School, Department of Biomedical Informatics, Cambridge, MA, 02115, US. ⁴Mount Sinai Hospital, New York, NY, 10029, US. ⁵Temple Health, Philadelphia, PA, 19140, US. ⁶VinBrain, Hanoi, Vietnam. ⁷Stanford University, Center for Artificial Intelligence in Medicine & Imaging, Stanford, CA, 94304, US. ⁸Copenhagen University Hospital, Department of Radiology, Copenhagen, Denmark. ⁹Stanford University, University Medical Line, Stanford, CA, 94305, US. ¹⁰Stanford University, Medical Center, Stanford, CA, 94305, US. ¹¹UC San Francisco, San Francisco, CA, 94143, US. ¹²Microsoft, Mountain View, CA, 94043, US. ¹³MIT and NBER, Department of Economics, Cambridge, MA, 02142, US. ¹⁴These authors contributed equally: Nikhil Agarwal, Pranav Rajpurkar, Tobias Salz. ✉e-mail: moehring@purdue.edu

access to clinical histories both affect radiologist performance. Our dataset provides predictions in two forms: a) as continuous probabilistic values of the likelihood a pathology is prevalent and b) as binary recommendations on treatment/follow-up. Typically, the treating physician, rather than radiologists, makes explicit treatment and follow-up decisions. We explained to radiologists that we would like them to make a recommendation to treat or follow-up on a pathology as if they were the treating physician. This question is only asked when relevant to a particular pathology.

The dataset contains additional information on the time spent by radiologists on each case and the values of two diagnostic standards defined by the researchers. The diagnostic standard is treated as the ground truth for the patient cases. We use this variable since a definitive diagnosis for most thoracic pathologies does not exist. These diagnostic standard variables are calculated by aggregating the assessments of five board-certified radiologists at Mt. Sinai Hospital and by estimating the leave-one-out mean of the probabilities reported by the radiologists in the experiment. An end-line survey provides information about the radiologists participating in the experiment.

To our knowledge, this dataset is the largest in the current literature on AI for radiology that can be used to compare human-AI collaboration with AI alone or humans alone. Unique features of the dataset include the large number of radiologists, disease labels and assessments for a comprehensive list of 104 pathologies, characteristics of the participants, and experimental variation in the availability of AI information. Past studies on AI for radiology have typically compared AI performance with radiologist performance, pitting humans against AI for a small set of diseases and with fewer experimental participants^{7–11}. In contrast, this dataset provides insight into human-AI collaboration, an increasingly promising paradigm for how AI models can be deployed in real-world radiology workflows⁶. This dataset also fills a gap in the field of AI-expert collaboration by providing a large experimental dataset in an important domain involving humans with significant expertise.

In the following sections, we summarize the data collected and discuss example uses, variables, and data collection.

Methods

Data Collection. The dataset was collected by researchers at the Massachusetts Institute of Technology (MIT) and Harvard University using a remote interface (see the Experiment Interface section for details). The MIT Committee on the Use of Humans as Experimental Subjects approved this experiment as an exempt study (protocol E-2953). Participants for this experiment were recruited from teleradiology companies and the VinMac healthcare system in Vietnam.

The dataset was generated through an experiment with 227 radiologists, of which 17% are US-based, and 20% have a degree from the US. The average radiologist in the dataset has 22 years of experience, and 38% of radiologists work in large clinical settings. Of the 162 radiologists who responded to the sex question, 62% are male. Close to 60% of the radiologists in the experiment had at least some previous experience working with AI tools.

The participating radiologists received links to the experiment platform, which included training information, an informed consent form, practice cases (only in experimental designs 1 and 2 which are described below), and the cases to be read during the study (see the Experiment Interface section for full details). The interface was designed to mimic the clinical setting for the participants. While we compensated the radiologists for participating in the experiment, we randomly provided monetary incentives for accuracy to a subset of radiologists in designs 1 and 3. Participants in design 2 did not receive any monetary incentives because of the nature of the recruitment agreement.

Historical Cases & AI. In each design, every radiologist views a subset of 324 historical cases procured from the Stanford Health Care System. These cases are manually reviewed for public release and contain a frontal X-ray and a summary of the patient's clinical history information¹². The X-ray images and clinical data can be obtained at (<https://stanfordaimi.azurewebsites.net/datasets/5194008e-61cf-4083-9896-3d4bd8bf8b0b>). After signing up and accepting the usage agreement, users can immediately download the data. The X-ray images can be linked to the Collab-CXR dataset⁵, as each patient ID in Collab-CXR corresponds to an X-ray image in the Stanford Health Care System dataset. This is an integer that has a corresponding X-ray image and indication. For instance, patient ID 1 corresponds to 1.jpg. The use of retrospective cases allows us to avoid ethical and other issues that would arise when experimenting in high-stakes settings. The AI support tool that we provide radiologists in the AI information environment comes from the CheXpert model which was developed by a team of researchers at Stanford University¹³. CheXpert is a deep learning algorithm trained on 224,316 chest radiographs of 65,240 patients and uses only the X-ray image to predict the prevalence of fourteen thoracic pathologies.

Diagnostic Standard. To establish a diagnostic standard for analyzing the quality of the radiologists' assessments, we aggregate data on the assessments of five board-certified radiologists at Mt. Sinai Hospital with at least ten years of experience and chest radiology as a sub-specialty. In radiology, a definitive diagnostic standard is generally not available and aggregating opinions from a panel of experts is commonly used to circumvent this issue. We also introduced a diagnostic standard constructed using a leave-one-out average of radiologists' assessments from the experiment. The reads used to calculate this diagnostic standard were from the treatment arm with clinical history but no AI assistance. We provide several potential aggregations of the panel of experts in addition to the individual decisions from this panel in the dataset.

Pathology Label Hierarchy. We designed and introduced a comprehensive system of labeling that balanced competing objectives of obtaining structured labels from a comprehensive set of pathologies, while not making the process too burdensome for experimental participants. Our extensive schema provides comprehensive

coverage of findings on chest X-rays. The labels span a wide range of pathologies including airspace opacities, lung nodules and masses, pleural abnormalities like effusions and pneumothorax, hilar and mediastinal findings, cardiovascular enlargement, degenerative chest wall changes, rib and spine fractures, foreign bodies, tubes and lines, and extra-thoracic findings. This broad scope matches the diverse diagnostic considerations facing radiologists interpreting chest X-rays.

A major strength of our schema is the detailed taxonomy within each category of findings. Airspace opacities are classified based on the presumed underlying etiology into entities such as pulmonary edema, pneumonia, ARDS, and atelectasis. Lung lesions are categorized into solid nodules, ground glass nodules, cavitary lesions, or masses. We delineate pneumothoraces into tension versus non-tension. This granularity allows for precise labeling that maps to specific diagnostic entities - critical for developing and evaluating high-performance AI algorithms.

Most existing chest X-ray labels for AI focus on a small subset of common pathologies like pneumonia, edema, and pneumothorax^{14,15}. For example, CheXpert only provides 14 labels covering common pathologies like pneumonia, edema, and effusions¹³. While important, these entities represent just a fraction of the diagnostic considerations on chest X-rays. In contrast, our schema contains over 100 distinct labels - far more extensive than existing schemes like CheXpert.

Our comprehensive labeling supports the development and assessment of AI that matches the breadth and specificity of radiologists' assessments. We provide greater granularity than CheXpert, separating out the enlargement of structures like the heart, hilum and mediastinum, while CheXpert has only a single "Enlarged Cardiomeastinum" label. Our schema also allows for multi-level specificity. We distinguish between general airspace opacity and specific etiologies like cardiogenic edema and pneumonia. This contrasts with other comprehensive schemas like that introduced in¹⁶, which categorize airspace opacity by focality rather than the underlying cause. Our approach maps labels directly to diagnostic entities, providing AI with the precise labels needed for high performance.

We recommend the following post-processing steps for category reorganization: Clavicle Fracture appears below the Rib Abnormality and Shoulder, and can be merged; In addition, the two occurrences of Spine Degenerative below Chest Wall Abnormality and Musculoskeletal Abnormality may be merged.

Experimental Designs. During the experiment, we collected radiologists' diagnoses on patient cases across four different information environments:

1. X-ray only
2. X-ray with clinical history information
3. X-ray with AI assistance
4. X-ray with AI assistance and clinical history information

Different experimental designs for collecting this information had various advantages and disadvantages. Therefore, we employed three distinct experimental designs (visualized in Fig. 1) to collect data on radiologist assessments under the four information environments. Within each design, radiologists were randomized into different sequences of information environments. The three designs together allow the study of both within and across subject variation.

Design 1. In this design, we assigned radiologists to a randomized sequence of the four information environments. Participants read 15 patient cases, 60 in total, under every information environment. These cases are read sequentially without a repeat encounter. At the beginning of the experiment, every radiologist reads eight practice cases. We recruited participants for design 1 from teleradiology firms.

Design 2. In this design, we asked radiologists to read cases in four experimental sessions separated by at least two weeks. Each radiologist was randomly assigned 60 cases to be read in each session. In each experimental session, every radiologist read their 60 cases under different information environments. This resulted in every radiologist reading the same 60 cases under each of the four possible information environments. Within each experimental session, each participant read 15 cases in each information environment. Information environments were presented in a random order in batches of five cases, with no cases repeated within a session. We recruited participants for design 2 from the VinMac healthcare system in Vietnam.

Design 3. In this design, participants read the same 50 cases both with and without access to AI assistance. We first asked the participants to read the 50 cases without access to AI assistance. Half of the cases were read without access to either AI assistance or clinical history and the other half were read without access to AI assistance but with access to clinical history. We randomized which cases received clinical history and whether or not a participant received the cases with clinical history first. We then asked the participants to read the same 50 cases with access to AI assistance. We only provided access to a patient's clinical history in the second read if the participant had access to clinical history in the first read. We recruited participants for design 3 from teleradiology firms.

Data Records and Dataset Structure

The Collab-CXR dataset is available at Open Science Framework (OSF) available at <https://osf.io/z7apq/5>. The dataset contains two compressed tab-delimited flat text files.

Each observation in the primary dataset refers to a radiologist-patient-pathology in a given experimental session. While there are no individual personal identifiers in the dataset, we have defined unique identifiers for

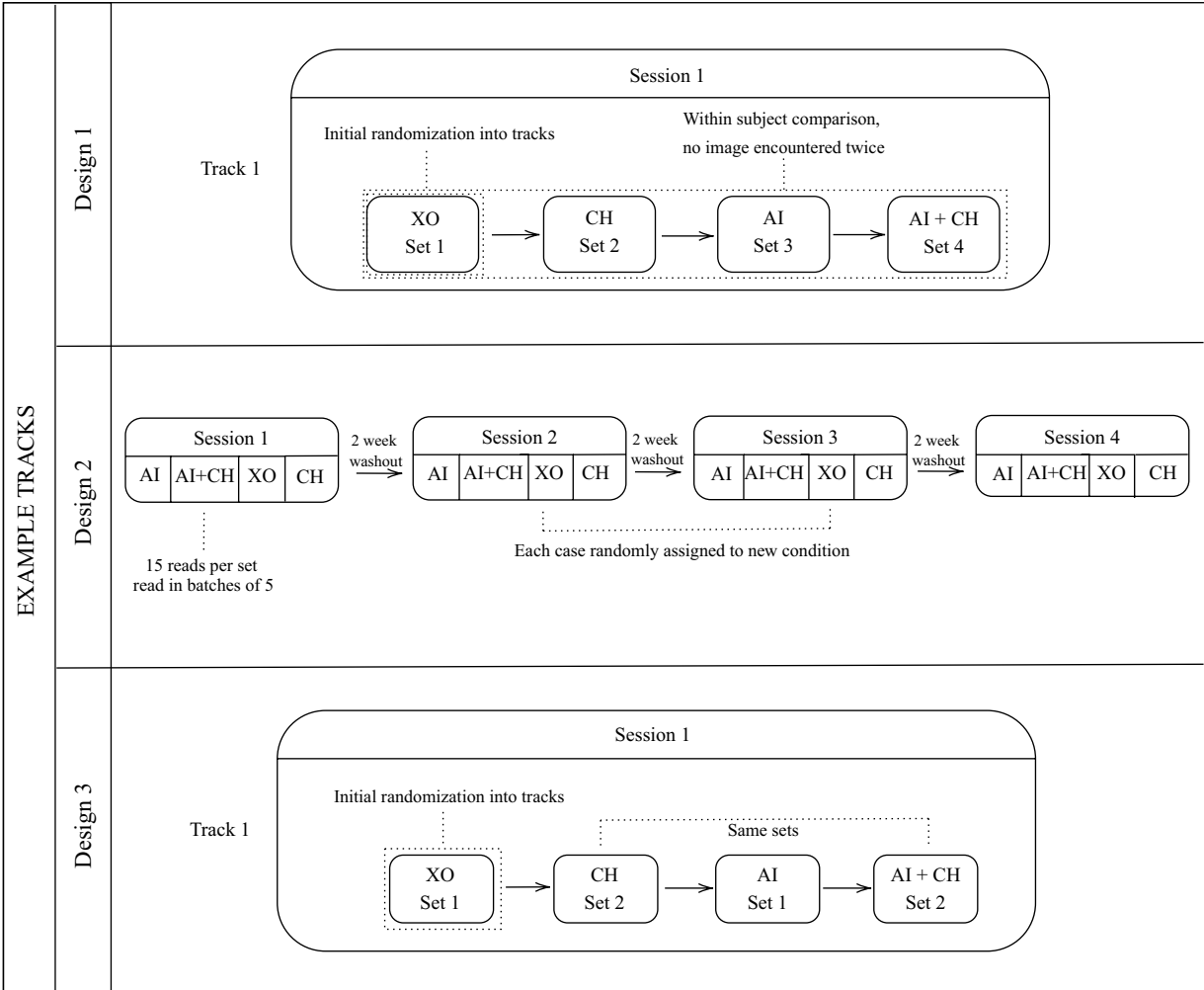


Fig. 1 Experimental designs. Note: This figure presents example tracks for the three separate research designs used in the experiment. The abbreviations represent the assigned information track. XO corresponds to x-ray only, CH corresponds to x-ray with clinical history information, AI represents x-ray with AI assistance, and AI + CH represents x-ray with AI assistance and clinical history information. In the first design, each radiologist is randomly assigned to one of 24 possible sequences of four information conditions. Within each condition, they evaluate 15 unique patient cases, and no radiologist sees the same patient more than once. In the second design, each radiologist evaluates a total of 60 patient cases across four distinct information environments. The experiment is divided into four sessions, with each session separated by a washout period of at least two weeks to minimize recall of previous assessments or AI assistance. In each session, radiologists assess 15 patients under each of the four environments, encountering each patient only once per session. In the third design, each radiologist reviews 50 patient cases, first without AI support and then with it. Clinical history is included for half of the cases, with its presence randomized to either the initial or latter portion of the session. The sets of patients evaluated with and without clinical history are distinct.

each radiologist and patient-case used in the experiment (variable names: “uid_clean” and “patient_id”). There is also information on the design under which the case was read (variable name: “design”) and the sequence in which each case was viewed by the radiologists within each experiment session (variable names: “round” and “experiment_session”). The dataset also includes both raw and processed diagnostic standard labels, as well as the individual reports by the five Mount Sinai diagnostic labelers. The dataset also includes elicited probabilities from experiment participants and AI for each radiologist-patient-pathology (variable names: “probability” and “alg_pred”). Table 1 summarizes the fields available in this dataset, and the data is stored on OSF in the file named “data_public.txt.gz”. In addition, we collect detailed click-stream data to log participant actions throughout the experiment. The click-stream data is described in Table 2 and is stored on OSF in the file named “click-stream_public.txt.gz”.

Code files are included in the https://github.com/mit-econ-ai/radiology_ai_data.

Post-experiment questionnaire. At the end of the experiment, radiologists complete a survey based on their experience with the AI support tool and clinical history in making their assessments. These endline survey responses are helpful in understanding the heterogeneity in radiologist qualities and how they use the various

Variable	Data Type	Range	Description
uid_clean	str		Radiologist-session identifier
experiment_id	str		Radiologist identifier
experiment_session	int	0, 1, 2, 3	Session number
design	dbl	1, 2, 3	Design arm
patient_id	int		Patient identifier
pathology	str		Pathology
round	long	$[1, 100] \in N$	Patient-case read sequence within radiologist-session (minimum value is 1)
treat	int	0, 1	Indicator if radiologist selected treat/follow-up.
			Missing for pathologies where it was deemed not relevant
treatment	str		Information environment under which case was read
			IN (NIN): Clinical history was (was not) provided,
			AI (NAI): AI assistance was (was not) provided,
			warmup: Practice image before the experiment started
level	int	0, 1, 2, 3	Position in pathology hierarchy (e.g. level 0 is top-level)
visible	int	0, 1	Indicator if pathology was visible when rad submitted case
probability	dbl	$[0, 1]$	Probability radiologist reported on interface slider
severity	str		Response to severity question, when relevant
size	str		Response to size question, when relevant
position	str		Response to position question, when relevant
active_time	dbl		Time spent actively working on case (in seconds)
raw_time	dbl		Total time spent on case (in seconds)
num_clicks	dbl		Number of clicks on a case
group_vietnam	int	0, 1	Indicator if the radiologist is from VinMac healthcare system, Vietnam
group_teleradiology	int	0, 1	Indicator if the radiologist is from a tele radiology company
group_pilot	int	0, 1	Indicator if the radiologist is from the experiment pilot
group_ground_truth	int	0, 1	Indicator if the radiologist is a ground truth radiologist
incentive_round	int	0, 1	Indicator if incentives for correct diagnosis were provided
total_num_rounds	long	$[65, 100]$	The total number of reads by one radiologist
chexbert_label	dbl	0, 1	ChexBert label for this case / pathology
alg_pred	dbl	$[0, 1]$	AI prediction
with_ai	byte	0, 1	Indicator if radiologist had access to AI
with_ch	byte	0, 1	Indicator if radiologist had access to CH
gt_average_XX	dbl	$[0, 1]$	Average of ground truth probabilities for group XX (us, vietnam, all, experiment)
gt_binary_simple_XX	long	0, 1	Binary ground truth based on simple average for group XX
gt_average_active_time_XX	dbl	(35, 344)	Average active time of the GT radiologists

Table 1. Variable List: Primary Experiment Data.

pieces of information. Their responses allow the analysis of the extent to which AI and clinical history affect the assessment, decision-making and effort exerted in the diagnosis process. This information also allows the exploration of questions regarding qualitative beliefs held by radiologists. The survey responses are also available on the OSF repository in the file “debrief_responses_osf.xlsx”.

Technical Validation

All validation checks are included in the appendix of the accompanying paper⁶. Below, we summarize some of the validation tests.

Validation of diagnostic standard quality. We present evidence that the diagnostic benchmark—based on readings from five board-certified radiologists at Mount Sinai, each of whom reviewed all 324 patient cases—is both high quality and robust to a range of analytic choices. We included a table with summary statistics for the diagnostic standard derived from the Mount Sinai radiologists, as well as for a leave-one-out internal diagnostic standard constructed using reads from the experimental arm that included clinical history but excluded AI assistance (Table C.3 in⁶) We also included additional summary statistics for the five labelers, such as the average time spent and click count (Table C.4 in⁶) We also report the average level of agreement between the diagnostic labels and the original interpretations provided by radiologists. (Table C.3 in⁶) These analyses indicate that the decisions of the five diagnostic standard labelers largely align with the clinical assessments made by the original radiologists, and we can statistically reject the null hypothesis that the average probability rating equals 0.5 at the

Variable	Data Type	Range	Description
uid_clean	str		Radiologist identifier
patient_id	str		Patient identifier
round	long	$[1, 100] \in N$	Patient-case read sequence within radiologist-session
			(minimum value is 1)
log_panzoom	json		List of dictionaries logging changes to x-ray images. Elements of each dictionary include
			x, y: Coordinates of user's pan (center of image is (0,0))
			zoom: Scale of the image (default is 1)
			image height / width: Height and width of the x-ray image
			utc time: UTC timestamp of the event
prboability_slider_change	json		List of dictionaries logging changes to probability sliders. Elements of each dictionary include
			pathology: Pathology being adjusted
			value: Probability being entered (ranges from 0-100)
			utc time: UTC timestamp of the event
log_xray_settings	json		List of dictionaries logging changes to the xray contrast and brightness. Elements of each dictionary include
			contrast: X-ray contrast
			brightness: X-ray brightness
			utc time: UTC timestamp of the event
log_click	json		List of dictionaries logging each click on the page. Elements of each dictionary include
			pageX, pageY: Coordinates of the click
			target closest: Closest element on the page clicked
			utc time: UTC timestamp of the event
set_to_zero	json		List of dictionaries logging when the radiologist clicked the "Set All to Highly Unlikely" button.
			Elements of the dictionary include
			paths set: List of pathologies set to zero
			utc time: UTC timestamp of the event
undo_set_to_zero	json		List of dictionaries logging when the radiologist clicked the undo button after "Set All to Highly Unlikely" button.
			Elements of the dictionary include
			paths undone: List of pathologies that were set to zero but are now undone
			utc time: UTC timestamp of the event
modal	json		List of dictionaries logging showing and hiding modals. Elements of the dictionary include
			modal: Name of the modal opened or closed
			event: show / hide to indicate if modal is opened or closed
			utc time: UTC timestamp of the event

Table 2. Variable List: Click-Stream Data.

5% significance level in the majority of cases. Finally, section C.5.2 in⁶ demonstrates the treatment effect results are consistent across methods of computing the diagnostic standard.

Validation of radiologist read quality. To benchmark the quality of AI predictions in our sample and the radiologists in our experiment, we compared our participant pool with the AI input using two performance measures (Figure 1 in⁶). The first measure is derived from the receiver operating characteristic (ROC) curve, which measures the trade-off between the false positive and the true positive rate of a classifier as the cutoff for classifying a case as positive or negative is varied. It uses only ordinal information about the AI. The second measure is the root mean-squared error (RMSE), which utilizes cardinal information about the AI prediction. We pool the data for top-level pathologies with AI for each radiologist's reports and for the AI's prediction.

Additional validation tests. We find that incentives do not play a significant role in diagnostic accuracy (Table C.20 in⁶). We also verify randomization occurred as expected through numerous balance and randomization tests (Figure C.9; Tables C.1, C.2 in⁶). For the second design, we also found that the washout strategy was effective. Radiologists' predictions do not shift towards the AI prediction if it was displayed in a prior session but do if it is displayed in the current session (Figure C.37 in⁶).

Code availability

Our [GitHub repository](#) includes code (Python 3 and Stata 17) for data preparation, technical validation, and summary statistics and post-experiment questionnaire.

Received: 4 November 2024; Accepted: 23 April 2025;

Published online: 03 May 2025

References

1. Ng, A. What artificial intelligence can and can't do right now. <https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now> Accessed: 2023-5-26 (2016).
2. Sezgin, E. Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. *DIGITAL HEALTH* **9**, <https://doi.org/10.1177/20552076231186520> (2023).
3. Norden, J. G. & Shah, N. R. What ai in health care can learn from the long road to autonomous vehicles. *Catalyst non-issue content* **3**, <https://doi.org/10.1056/CAT.21.0458> (2022).
4. Langlotz, C. P. Will artificial intelligence replace radiologists? *Radiology: Artificial Intelligence* **1**, e190058, <https://doi.org/10.1148/ryai.2019190058> (2019).
5. Moehring, A. *et al.* Collab-cxr: A dataset for understanding radiologist - artificial intelligence collaboration, <https://doi.org/10.17605/OSF.IO/Z7APQ> (2023). OSF, CC-BY 4.0 License. Last updated: 2025-04-03.
6. Agarwal, N., Moehring, A., Rajpurkar, P. & Salz, T. Combining human expertise with artificial intelligence: Experimental evidence from radiology, <https://doi.org/10.3386/w31422> (2023).
7. Huang, J. *et al.* Generative artificial intelligence for chest radiograph interpretation in the emergency department. *JAMA Network Open* **6**, e2336100–e2336100, <https://doi.org/10.1001/jamanetworkopen.2023.36100> (2023).
8. Wu, J. T. *et al.* Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents. *JAMA Network Open* **3**, e2022779–e2022779, <https://doi.org/10.1001/jamanetworkopen.2020.22779> (2020).
9. Majkowska, A. *et al.* Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* **294**, 421–431, <https://doi.org/10.1148/radiol.2019191293> (2020).
10. Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the cheXnext algorithm to practicing radiologists. *PLOS Medicine* **15**, 1–17, <https://doi.org/10.1371/journal.pmed.1002686> (2018).
11. Tiu, E. *et al.* Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering* **6**, 1399–1406, <https://doi.org/10.1038/s41551-022-00936-9> (2022).
12. Stanford Center for Artificial Intelligence in Medicine and Imaging (AIMI). Chexmultimodal: A multimodal dataset of chest x-rays and clinical history from stanford university hospital, <https://stanfordaimi.azurewebsites.net/datasets/5194008e-61cf-4083-9896-3d4bd8bf8b0b> (2024).
13. Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence* **33**, 590–597, <https://doi.org/10.1609/aaai.v33i01.3301590> (2019).
14. Johnson, A. E. *et al.* Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* **6**, 317, <https://doi.org/10.1038/s41597-019-0322-0> (2019).
15. Wang, X. *et al.* Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* <https://doi.org/10.48550/arXiv.1705.02315> (2017).
16. Seah, J. C. *et al.* Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet Digital Health* **3**, e496–e506, [https://doi.org/10.1016/S2589-7500\(21\)00106-0](https://doi.org/10.1016/S2589-7500(21)00106-0) (2021).

Acknowledgements

The authors acknowledge support from the Alfred P. Sloan Foundation (2022-17182), JPAL Healthcare Delivery Initiative, and MIT SHASS.

Author contributions

A.M., N.A., T.S., and P.R. were responsible for conceptualization, management, and experimental design. A.M., M.K., R.H., O.B., N.A., T.S., and P.R. contributed to the analysis and writing. A.J., C.E., D.M., M.C., E.D., Y.G., T.B., and S.T. contributed by testing the experimental interface and reading patient cases. A.P., C.L., M.L. contributed to the clinical aspects and interpretation of the results. A.P. and M.L. consulted on the design of the experimental interface.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-025-05054-0>.

Correspondence and requests for materials should be addressed to A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025