Learning from Machine Learning: Evaluating How Contextual Interference Impacts the Learning of a Radiology Task in Machines and Radiologists

Brendan T. Crabb, BS[1], Megan K Mills, MD[1], Laruen Williams, PhD[2], Angela Presson, PhD[3], Chong Zhang, MS[3], Satoshi Minoshima, MD, PhD[1]

SUMMARY

In this randomized controlled trial, interleaved training schedules did not have a statistically significant effect on the learning of a radiology task for a cohort of radiology residents or machine learners.

OBJECTIVE

Utilizing the theory of contextual interference, this study sought to assess the impact of interleaved (non-blocked) training schedules on the learning of mass detection on mammograms and nodule detection on chest radiographs by radiology residents and a convolutional neural network (CNN).

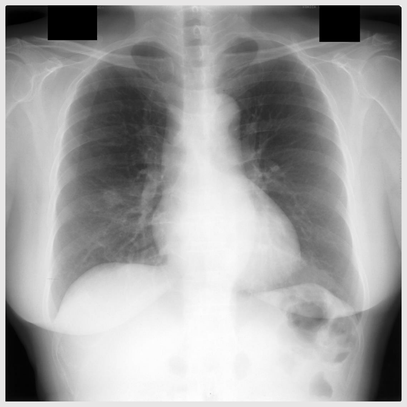
METHODS

Human Learners

With IRB-exemption, fifteen junior radiology residents with experience ranging from PGY-1 to PGY-3 were enrolled and randomized to an interleaved (non-blocked) or control (blocked) training schedule **(Figure 1)**. Chest radiographs were collected from the Standard Digital Image Database from the Japanese Society of Radiological Technology (JSRT) and mammograms were collected from the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM). After assessing baseline performance with a pre-test of 40 radiographs, participants were taught to identify masses on mammograms and nodules on chest radiographs according to their assigned training schedule. Participants completed three sessions of radiographic image viewing and education via an internet-based software platform. Following training, performance characteristics including accuracy, sensitivity, specificity, and response time were measured on 40 novel radiographs. Training schedule effect was analyzed using mixed effects log binomial regression models and Wald tests.

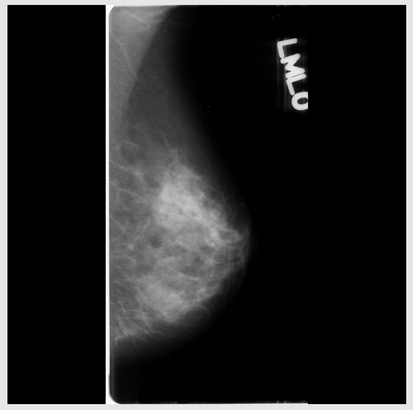
Machine Learners

Chest radiographs and mammograms, with and without nodules and masses, were collected from the JSRT, CBIS-DDMS, and NIH ChestX-ray14 datasets, producing a final dataset with 2,880 chest radiographs and 2,880 mammograms. The data was split at the patient-level into a training set (5,680 images) and testing set (80 images). A ResNet50-based multi-task CNN architecture with base-weights from ImageNet **(Figure 2)** was then trained to perform image classification with either an interleaved (non-blocked) or control (blocked) schedule. The interleaved schedule alternated tasks at the batch level, while blocked training was performed at the epoch level. For both schedules, 5-fold cross-validation on the training set was utilized. Post-training performance was evaluated on the same 40 radiographs as the human learners. Training schedule effect was analyzed using mixed effects log binomial regression models and Wald tests.



(WO)MAN VS MACHINE

Can you find the nodule and mass?



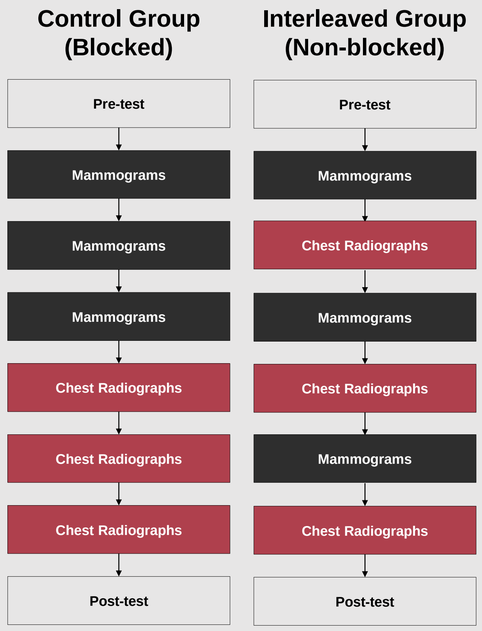


Figure 1: Representative example of an interleaved training schedule and a blocked training schedule (control group). Participants and machine learners were randomly assigned to one of these two schedules.

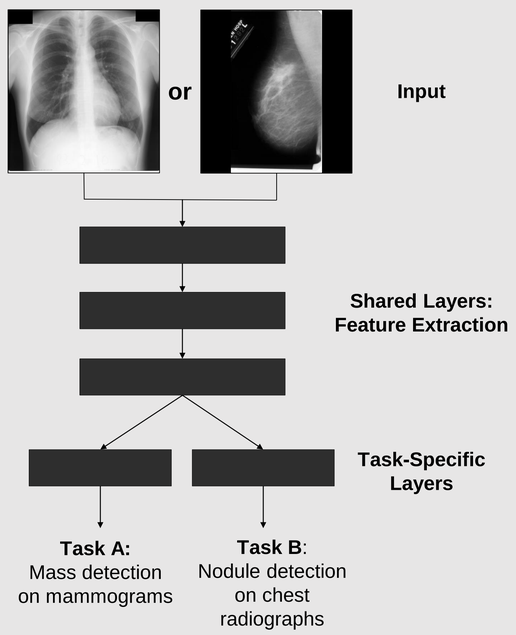


Figure 2: Schematic of a multi-task neural network with a shared trunk architecture, where both radiology tasks share common neural network layers for feature extraction.

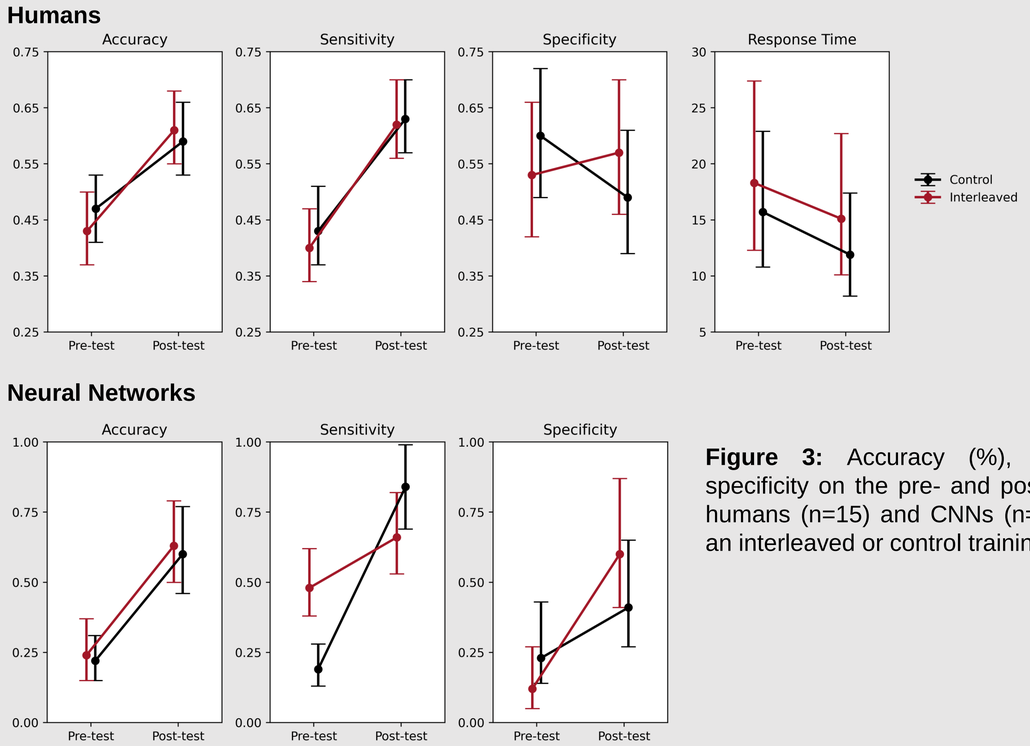


Figure 3: Accuracy (%), sensitivity, and specificity on the pre- and post-test exams for humans (n=15) and CNNs (n=10) trained with an interleaved or control training schedule.

RESULTS

Fifteen residents completed the training (7 interleaved, 8 control). In both groups, training increased overall accuracy (0.45 to 0.60, p<0.001) and sensitivity (0.42 to 0.63, p<0.001), but not specificity (0.57 to 0.53, p=0.49), as shown in **Figure 3**. For residents, training schedule had no significant effect on the ratio of pre- to post-test sensitivity or specificity (sensitivity RR=1.07 (95% CI 0.82 to 1.39), p=0.61; specificity RR=1.32 (95% CI 0.87 to 1.99), p=0.19).

Similarly, CNNs trained with both interleaved and control schedules showed improvements in accuracy (0.29 to 0.69, p<0.001), sensitivity (0.34 to 0.75, p<0.001) and specificity (0.17 to 0.51, p<0.001), as shown in **Figure 3**. However, no significant differences for post-test accuracy, sensitivity, and specificity were found between interleaved and control training schedules (accuracy 0.63 vs. 0.60, p=0.98; sensitivity 0.66 vs 0.84, p=0.75; specificity 0.60 vs 0.41, p=0.93).

DISCUSSION

In this randomized controlled trial, no difference in performance on radiographic interpretation was observed between blocked and interleaved training schedules for radiology residents or machine learners.

For both human and machine learners, there was a trend towards improved post-test accuracy using an interleaved training schedule. The lack of statistical significance in this scenario may be secondary to the limited sample size. Additional studies are necessary to determine the full impact of contextual interference on the learning of radiological tasks.

ACKNOWLEDGEMENTS

This investigation was supported by the Association of Residency Program Directors Jerome Arndt grant and by the University of Utah Study Design and Biostatistics Center, with funding in part from the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1TR002538.

AUTHOR AFFILIATIONS

1. Department of Radiology and Imaging Sciences, University of Utah School of Medicine, Salt Lake City, UT
2. Department of Psychology, University of California San Diego, La Jolla, CA
3. Department of Epidemiology, University of Utah School of Medicine, Salt Lake City, UT