Artificial Intelligence in Breast Ultrasound: Moving from Standalone Performance to the Physician-system Interface

Lev Barinov, Princeton University, Rutgers Robert Wood Johnson Medical School, Koios Medical; Ajit Jairaj; Stephen Seymour; Diane Quigley

Introduction
Ultrasound (US) is an invaluable tool for evaluating dense breast tissue, with comparable sensitivity to digital mammography (DM) and improved detection of invasive and node-negative cancers. This comes at the cost of lower PPV3 and specificity [1]. In a recent study, an AI based Decision Support (DS) system was shown to exceed radiologist performance when analyzing US data (34-55% reduction in benign biopsies and a PPV3 increase of 7-20%)[2]. Studies evaluating the impact of AI and DS systems should measure performance in the context of clinical workflows. The goal of this study is to evaluate workflow schemas for DS integration and their effects on diagnostic accuracy and efficiency.

Hypothesis

Methods
500 anonymized cases (368 (74%) benign and 132 (26%) malignant) were retrospectively selected for inclusion. There was a benign BI-RADS breakdown of 11%, 19%, 70%, and 0%, and a malignant breakdown of 0%, 1%, 54%, and 45% for BI-RADS categories 2-5. Three board-certified readers with 20+, 10+, and 3 years of breast reading experience were included.

Two reading and workflow methodologies are evaluated: sequential and independent. During the study, readers examine two orthogonal views of each lesion and assess BI-RADS category and Likelihood of Malignancy. Phase-1 of the study examines Sequential Reads (SR). Readers are asked to perform a Control Read (CR) and make an initial assessment. They are then presented with a DS recommendation and reassess. A 4-week washout period is then conducted to prevent bias via lesion-recall. Phase-2 examines Independent Reads (IR), where readers are presented with DS recommendations from the outset. 50 additional Control Reads (CR) are presented with no DS recommendation at all, to assess intra-reader variability. Reader performance is measured using AUC analysis. Inter-reader and intra-reader variability is measured using the Kendall Tau-b statistic.

Figure
**Results**

As per Figure 1, the system achieved an AUC of 0.8648 [0.8345-0.8893]. Each of the three readers attained CR performance of 0.7618 [0.7244-0.7934], 0.7543 [0.7197-0.7887], and 0.7325 [0.6897-0.7689], respectively. Their SR AUCs were 0.7935 [0.7567-0.8229], 0.7674 [0.7327-0.8001], and 0.7859 [0.7527-0.8174] and IR AUCs were 0.8213 [0.7861-0.8516], 0.8305 [0.7982-0.8594], and 0.7988 [0.7632-0.8310].

Comparison of CR and SR reads, by reader, yielded Z-scores of 1.1858, 0.52313, and 1.9332. Z-scores comparing CR and IR were 2.1896, 3.1644, and 2.41. Z-scores comparing CR to the system were 4.1144, 4.6984, and 5.5813. Intra-reader variability was measured, via Kendall Tau-b, at 0.597, 0.595, and 0.529. Average inter-reader variability was 0.474 and 0.698 between CR and IR readers, respectively, with all pairwise values showing statistical significance.

**Conclusion**

The significant performance gap between the reading methodologies makes it apparent that standalone performance metrics are insufficient for evaluating AI and DS system efficacy and impact. These results raise the prospect of future work assessing the most effective means of integrating high performance automated assessments into clinical workflows.

**Statement of Impact**

This study's findings demonstrate that additional concerns and evaluation criteria are necessary when integrating AI and machine learning based systems into clinical practice.

**Keywords**

machine learning, decision support, artificial intelligence, breast ultrasound.

**References**:

1. Wendie A. Berg, Andriy I. Bandos, Ellen B. Mendelson, Daniel Lehrer, Roberta A. Jong, Etta D. Pisano; Ultrasound as the Primary Screening Test for Breast Cancer: Analysis From ACRIN 6666, *JNCI: Journal of the National Cancer Institute*, Volume 108, Issue 4, 1 April 2016, djv367, [https://doi.org/10.1093/jnci/djv367](https://doi.org/10.1093/jnci/djv367)