CONCORDANCE OF MAMMOGRAPHIC STUDIES BETWEEN ARTIFICIAL INTELLIGENCE AND EXPERT BREAST RADIOLOGISTS

R. Limon[1,4], C. Antelo[3], H. Adorno [1], R. Ticona [3], W. Cortez[1], E. Claros[5], N. Espinoza [2, 3], C. Henostrosa[4], M. Gianella[1], D. Ramirez [2, 3], C. Sitic[2,3], E. Rivero [7], D. La Fuente [1]. C. Oporto [1,4], O. Niño de Guzman [9], C. Rocha [1], C. Salcedo [1], A. Condarco [1], E. Cardoso [1], H Gomez[6], N. Valdivieso[6], S. Neciosup[8]

1.Clinica De Las Americas Santa Cruz-Bolivia; 2. Perinat, Santa Cruz-Bolivia; 3. Instituto Oncológico Oriente Boliviano Santa Cruz- Bolivia; 4. OncoBolivia Santa Cruz - Bolivia; 5.Oncos Santa Cruz-Bolivia; 6. Oncosalud Auna Ideas, Lima-Perú, 7. Centro Medico FOIANINI Santa Cruz – Bolivia 8. Clinica Internacional Lima- Peru, 9. Centro de Oncologia Ginecologica Niño de Guzman

PS1-13-30

Background

Various artificial intelligence (AI) programs have been specifically designed and trained for breast cancer screening and prediction, and their use is steadily increasing. However, tools such as ChatGPT—an AI not originally developed for imaging analysis—are being increasingly employed for this purpose, particularly by non-medical personnel. This situation may generate mistrust and uncertainty among patients when comparing AI-generated reports with those of medical specialists. Therefore, the proper use of these tools requires awareness of their intended scope and limitations. The purpouse of this study is to evaluate the agreement and consistency of ChatGPT when assising BIRADS (Breast Imaging Reporting and Data System) and ACR (Breast Density American College of Radiology) categories in mammographic studies and to compare them with radiologist who are experts in breast imaging studies.

Methods

Random mammograms performed in August 2025 on asymptomatic women undergoing preventive screening were included. All studies were interpreted and reported by expert breast radiologists. For AI evaluation, images were uploaded using an iPhone 16 Pro Max into ChatGPT, which generated a structured mammography report. Concordance was assessed across two parameters: BI-RADS classification and ACR breast density Two primary variables were recorded for each case: BI-RADS suspiciousness category, and ACR breast density class (A–D). The radiologist performed the standard clinical interpretation; afterwards, the images and descriptive findings were processed by ChatGPT-5 using a standardized protocol. Agreement between evaluators was assessed using: Confusion matrices for BI-RADS and ACR. Unweighted Cohen's Kappa for exact agreement. Linear and quadratic weighted Kappa for ordinal-scaled variables. Observed and expected agreement proportions were calculated.

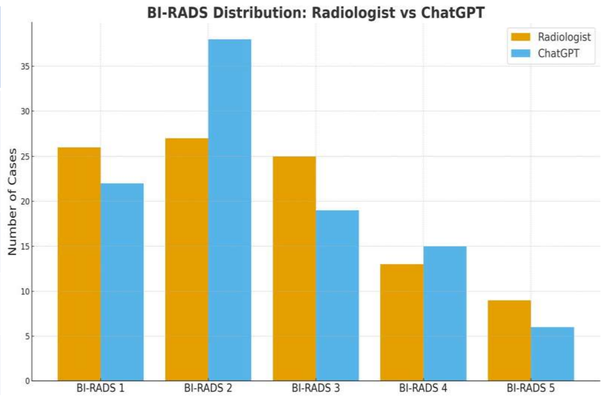
Results

Radiologists / ChaGPT	BIRADS 1	BIRADS 2	BIRADS 3	BIRADS 4	BIRADS 5	TOTAL
BIRADS 1	17	4	3	2	0	26
BIRADS 2	2	23	1	1	0	27
BIRADS 3	3	5	12	3	2	25
BIRADS 4	0	6	1	5	1	13
BIRADS 5	0	0	2	4	3	9
TOTAL	22	38	19	15	6	100

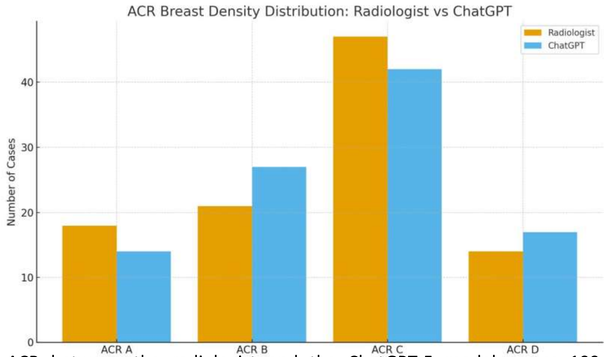
Fig. 1 Radiologists * ChatGPT crosstabulation BIRADS

Radiologists / ChaGPT	ACR A	ACR B	ACR C	ACR D	TOTAL
ACR A	8	4	3	3	18
ACR B	2	15	2	2	21
ACR C	3	5	35	4	47
ACR D	1	3	2	8	14
TOTAL	14	27	42	17	100

Fig. 2 Radiologists * ChatGPT crosstabulation ACR



BI-RADS categories assigned by the radiologist and the ChatGPT-5 evaluators show similar classification patterns, with the highest frequency in BI-RADS 2 and the lowest in BI-RADS 5



ACR between the radiologist and the ChatGPT-5 model across 100 mammography studies. Evaluations show similar patterns, with the highest frequency in ACR C.

BI-RADS agreement: In the 100 cases evaluated, exact agreement between the radiologist and ChatGPT-5 for BI-RADS classification was 60%. The unweighted Cohen's Kappa was 0.48, indicating moderate agreement.

Given the ordinal structure of BI-RADS, weighted measures were calculated:

- Linear weighted kappa: 0.54
- Quadratic weighted kappa: 0.63

These values represent moderate to substantial agreement.

Most discrepancies occurred between adjacent categories (BI-RADS 2–3 and 4–5), a clinically expected pattern.

Breast density ACR agreement:

ACR density categories (A–D), exact agreement was 66%. The unweighted Cohen's Kappa was 0.51, reflecting moderate agreement. Weighted analysis yielded:

- Linear weighted kappa: 0.63
- Quadratic weighted kappa: 0.71

indicating substantial agreement between the radiologist and ChatGPT-5 Most mismatches appeared between neighboring density levels, particularly ACR B–C and ACR C–D.

Category distribution: Both evaluators demonstrated similar classification patterns for BI-RADS and ACR density, with the highest frequency in BI-RADS 2 and ACR C. Bar charts showed minimal deviation in category distribution across evaluators.

Conclusions

ChatGPT-5 showed moderate to substantial agreement with the radiologist for both BI-RADS and ACR. Most discrepancies occurred between adjacent categories, with minimal clinical impact. Category distribution was highly similar between evaluators. Findings support the model's potential role as an assistive tool in mammography interpretation.

Contact

RONALD SERGIO LIMON TELLEZ
OncoBolivia
Email: r_limon85@Hotmail.com
Phone: + 591-77372036

References

1. Nigar Erkoc, Elif Hazal Karlı, Emre Tunay, Ahmet Tan Cimilli, Concordance between artificial intelligence and radiologists in BIRADS classification of breast ultrasound: A study using ChatGPT-4o, Clinical Imaging,Volume 128, 2025, 110636, ISSN 0899-7071
2. P. Sun et al. Preliminary experiments on interpretable ChatGPT-assised diagnosis for breast ultrasound radiologists. Quan Iamging Med Surg 2024
3. Y.C Gunes et al. Evaluating text and visual diagnostic capabilities of large language models on question related to the breast imaging reporting and data system %th edition

This presentation is the intellectual property of the author/presenter. Contact them at <<your email address>> for permission to reprint and/or distribute.

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com