Reliability of the BI-RADS Final Assessment Categories and Management Recommendations in a Telemammography Context

Antonio J. Salazar, PhD[a], Javier A. Romero, MD, MSc[b], Oscar A. Bernal, MD, PhD[c], Angela P. Moreno, MD, MSc[b], Sofía C. Velasco, MD, MSc[b]

Abstract

Purpose: The aim of this study was to evaluate the intradevice and interdevice reliability of four alternatives for telemammography—computed radiography, printed film, a film digitizer, and a digital camera—in terms of interpretation agreement when using the BI-RADS® lexicon.

Methods: The ethics committee of the authors' institution approved this retrospective study. A factorial design with repeated measures with 1,960 interpretations was used (70 patients, seven radiologists, and four devices). Reliability was evaluated using the κ coefficient for intradevice and interdevice agreement on malignancy classification and on BI-RADS final assessment category.

Results: Agreement on malignancy classification was higher than agreement for BI-RADS final assessment category. Interdevice agreement on malignancy classification between the film digitizer and computed radiography was ranked as almost perfect ($P < .001$), whereas interdevice agreement for the other alternatives was ranked as substantial ($P < .001$), with observed agreement ranging from 85% to 91% and κ values ranging from 0.70 to 0.81. Interdevice agreement on BI-RADS final assessment category was ranked as substantial or moderate ($P < .001$), with observed agreement ranging from 64% to 77% and κ values ranging from 0.52 to 0.69. Interdevice agreement was higher than intradevice agreement.

Conclusions: The results of this study show very high interdevice agreement, especially for management recommendations derived from malignancy classification, which is one of the most important outcomes in screening programs. This study provides evidence to suggest the interchangeability of the devices evaluated, thereby enabling the provision of low-cost medical imaging services to underserved populations.

Key Words: BI-RADS, breast cancer, film digitizer, mammography, reliability, telemammography

J Am Coll Radiol 2017;14:686-692.

INTRODUCTION

Variability in radiologists' interpretations may reduce the accuracy of mammography in the early detection of breast cancer [1]. To standardize the reporting of findings in different imaging modalities, the ACR developed the BI-RADS® atlas [2]. Several studies have evaluated the accuracy of this system compared with traditional mammography [3,4]. Other studies have evaluated the accuracy of mammography, ultrasound, and physical examination, compared with biopsy findings, when using the BI-RADS lexicon [5,6].

Assessments regarding the BI-RADS atlas usually concern feature analysis (eg, breast density, lesion type, mass borders, mass density, mass shape, microcalcification morphology, microcalcification distribution [3]), as well as assessments of management recommendations (eg, routine mammographic screening, short-interval follow-up, tissue diagnosis) [3-5]. According to these studies, results obtained using the BI-RADS categories have been found to be useful in differentiating between benign and malignant breast lesions [7].

[a]Electrophysiology and Telemedicine Laboratory, University of Los Andes, Bogotá, Colombia.

[b]Department of Diagnostic Imaging, Fundación Santa Fe de Bogotá University Hospital, Bogotá, Colombia.

[c]School of Government, University of Los Andes, Bogotá, Colombia.

Corresponding author and reprints: Antonio J. Salazar, PhD, Electrophysiology and Telemedicine Laboratory, University of Los Andes, Carrera 1 Este No. 19A-40, Bogotá, Colombia; e-mail: ant-sala@uniandes.edu.co.

The authors have no conflicts of interest related to the material discussed in this article.

1546-1440/16 ■ http://dx.doi.org/10.1016/j.jacr.2016.08.004

In underserved areas, telemedicine, using digital images, may provide a cost-effective solution for screening mammography programs. Previous studies have reported no significant differences between screen-film mammography and digital modalities, such as computed radiography (CR) and full-field digital mammography [8,9]. Nevertheless, these technologies are still unaffordable for the vulnerable populations of Colombia, especially in the Amazonian jungles where only conventional screen-film mammography is available, if any radiologic services exist at all [10]. In previous studies, we evaluated the validity of low-cost telemammography configurations, in terms of sensitivity, specificity, and receiver operating characteristic curves [11]. In this study, the aim was to assess reliability among different solutions for telemammography, such as film digitizers, digital cameras, printed film, and CR, in terms of interpretation agreement, over interpretation results based on the BI-RADS lexicon.

METHODS

The ethics committee of our institution approved this retrospective study, and informed consent was not required. This study applied a design with repeated measures, using 70 patients, seven radiologists, the reference images (ie, CR) and three derived images, for a total of 1,960 readings.

To perform validity assessments, sensitivity, specificity, and positive predictive value are usually evaluated. In contrast, in the context of this study, reliability was the reproducibility or agreement in measurements of the variables for each case when rated by different observers (ie, intradevice reliability) or when rated by each observer using different treatments (ie, interdevice reliability).

The Reference Standard

The actual state of the mammograms enabled us to determine the distribution of the sample. The standard for positive cases was a malignant lesion confirmed by biopsy within 2 years of the initial mammographic screening [8,9,12]. Negative cases were defined as those without any lesions confirmed by biopsy or those with normal results on follow-up mammography for 2 years. Two radiologists with more than 10 years of experience in reading mammograms, with access to the clinical histories of the patients, established the reference standard.

Study Sample and Readers

At most rural health centers in our country, there are no mammographic services [10]. As such, there are no mammograms available for use in a retrospective study. For these reasons, this study was undertaken using computed radiographic screening mammograms from our hospital, which sees many patients from such underserved areas of our country.

Random screening mammograms from asymptomatic patients who attended mammographic screening at the Fundación Santa Fe de Bogotá University Hospital, performed over 2 years, were included in the study sample. Of these patients, 50% had management recommendations for tissue diagnosis and 50% for follow-up. To be included, each case was required to include the four standard mammographic views. Cases of tomosynthesis or large lesions were excluded.

The sample size was determined in our previous studies to be 70 cases, with an approximately 1:1 ratio of malignant and benign cases. The patients ranged in age from 41 to 84 years, with a mean age of 62.1 $\pm$ 11.5 years. There were 57 cases with calcifications, 26 with masses, 35 with asymmetries, and 11 with architectural distortions and associated features. Four patients with prostheses were also included in the sample.

The distribution of cases according to BI-RADS final assessment category was as follows: 18 in category 2, 19 in category 3, 6 in category 4A, 14 in category 4B, 3 in category 4C, and 10 in category 5. In terms of malignancy, there were 33 patients with malignant lesions and 37 patients with benign lesions or normal results. Detailed classification of the cases in the sample, and their distribution according to the BI-RADS final assessment categories are presented in supplemental tables (S1 and S2).

Seven radiologists from Fundación Santa Fe de Bogotá University Hospital (including four with more than 10 years of experience in mammography who were well trained in the BI-RADS lexicon and three radiologists with more than 1 year of experience who were trained in the BI-RADS lexicon for the purposes of this study) served as the observers.

Observed Variables by Radiologists

Data collection was performed using a database and a digital form that was integrated into the image-viewing software. At each interpretation, the radiologist selected the level of confidence in the presence of masses, calcifications, asymmetries, and architectural distortions and associated features. The radiologists were asked to classify the lesion features, such as mass borders, mass density, mass shape, microcalcification morphology, microcalcification distribution, asymmetric density, and architectural distortion. Additionally, the radiologist classified