Original Investigation

The Reproducibility of Changes in Diagnostic Figures of Merit Across Laboratory and Clinical Imaging Reader Studies

Frank W. Samuelson, PhD, Craig K. Abbey, PhD

Rationale and Objectives: In this paper we examine which comparisons of reading performance between diagnostic imaging systems made in controlled retrospective laboratory studies may be representative of what we observe in later clinical studies. The change in a meaningful diagnostic figure of merit between two diagnostic modalities should be qualitatively or quantitatively comparable across all kinds of studies.

Materials and Methods: In this meta-study we examine the reproducibility of relative measures of sensitivity, false positive fraction (FPF), area under the receiver operating characteristic (ROC) curve, and expected utility across laboratory and observational clinical studies for several different breast imaging modalities, including screen film mammography, digital mammography, breast tomosynthesis, and ultrasound.

Results: Across studies of all types, the changes in the FPFs yielded very small probabilities of having a common mean value. The probabilities of relative sensitivity being the same across ultrasound and tomosynthesis studies were low. No evidence was found for different mean values of relative area under the ROC curve or relative expected utility within any of the study sets.

Conclusion: The comparison demonstrates that the ratios of areas under the ROC curve and expected utilities are reproducible across laboratory and clinical studies, whereas sensitivity and FPF are not.

Key Words: Sensitivity; specificity; AUC; reproducibility.

Published by Elsevier Inc. on behalf of The Association of University Radiologists.

INTRODUCTION

Studies of reader performance play an important role in evaluations of the effectiveness of new diagnostic imaging devices and methodologies (1–5). These studies are particularly common in applications where a diagnostic task can be simplified to a binary choice, as in breast cancer screening where the goal is to identify abnormalities for subsequent analysis. From this perspective, the screening examination sorts the cases into "negative" and "positive" categories. Binary tasks are commonly characterized by a receiver operating characteristic (ROC) curve, which plots true positive fraction (TPF; the fraction of disease cases correctly labeled positive) as a function of the false positive fraction (FPF; the fraction of nondisease cases incorrectly labeled as positive). The ROC curve serves as the underlying framework from which various figures of merit (FOMs) of imaging performance can be defined.

Before using a new diagnostic imaging modality in clinical practice, a laboratory reader study of the modality may be performed to determine its diagnostic effectiveness for regulatory purposes or to demonstrate its capabilities for the medical community. Often this study will compare the performance of the new modality in one study arm to the standard of care in a control arm (1,4,6–8). These preclinical reader studies differ from clinical trials in a number of ways that make the studies far less time-consuming and costly. They generally require a much smaller sample of patient examinations, which limits the potential for side effects of imaging, including exposure to ionizing radiation or intravenous contrast agents. In imaging tasks that have low disease prevalence, for example, asymptomatic breast-cancer screening, the patient sample is usually enriched with cases of disease through various possible sampling strategies (9). Other differences may include limited access to additional clinical data (prior examinations, patient symptoms, or history), a different reporting format that may include nonstandard measures such as a probability of

Acad Radiol 2017; ■:■■–■■

From the U.S. Food and Drug Administration, 10903 New Hampshire Ave., Building 62, Room 3102, Silver Spring, MD 20993-0002 (F.W.S.); Department of Psychological and Brain Sciences, University of California, Santa Barbara, California (C.K.A.). Received January 9, 2017; revised April 28, 2017; accepted May 1, 2017. **Address correspondence to:** F.W.S. e-mail: frank.samuelson@fda.hhs.gov

Published by Elsevier Inc. on behalf of The Association of University Radiologists.
http://dx.doi.org/10.1016/j.acra.2017.05.007

malignancy rating, and use of retrospective data that results in a lack of direct consequences for patient management.

After initial acceptance of an imaging modality, clinical observation studies are often reported as a way to provide an assessment of the new modality in practice. These studies typically compare the new modality to the preexisting standard in a clinical setting with a clinical population and with patient management decisions based on outcomes from imaging. Observational studies are instrumental for determining the success of a new imaging modality in the health-care environment, for making reimbursement decisions, and potentially for defining a new standard of care.

The literature on preclinical laboratory studies includes an extended debate over what should be the appropriate FOM for quantifying diagnostic performance of imaging systems (2,5,10–16). Proposed FOMs include TPF and FPF (17–19), area under the ROC curve (AUC) (20–22), and utility measures (23–27) among others (28–31). TPF and FPF are considered criterion-dependent measures because they result from a perceived signal exceeding a critical value in classical signal detection models. They are understood in these models as an operating point on an ROC curve. In contrast, AUC is considered criterion-free because it does not depend on any one operating point. We also note that there are other approaches to defining imaging performance based on assessing localization accuracy (32–34) with a similar debate over FOMs (35–37) that this work will not address.

Much of the debate has focused on how an FOM may be interpreted as a measure of performance, or what advantages an FOM may have in experimental design such as statistical power. A consideration that has received far less attention is the issue of reproducibility. A most fundamental feature of any FOM is its ability to measure differences between experimental conditions reproducibly. If an FOM cannot be reproduced quantitatively, or at least qualitatively, across scientific studies, then it is not useful. Reproducibility is the focus of this work.

Much recent literature has been devoted to the importance of the reproducibility of scientific results (38), particularly in medicine. In this work we examine the reproducibility of different FOMs across initial laboratory studies and how well those FOMs translate to the subsequent clinical observational studies. Specifically, we are interested in which FOMs demonstrate qualitative or quantitative reproducibility of measured changes between the experimental and control arms of the studies. In this paper we use a broad definition of the term reproducibility to mean how well an FOM is reproduced across both controlled and observational studies with different designs and settings.

We compare several published laboratory studies on screen film mammography (SFM), full-field digital mammography (FFDM), breast ultrasound (US) imaging, and digital breast tomosynthesis (DBT) with the published observational studies that followed. These studies examine imaging technologies used to screen patients for breast cancer. We focus on imaging modalities related to breast cancer screening because these have received considerable interest over time, and hence there are more published data available.

There are a number of technical challenges when computing the FOMs from the data reported in observational studies. A primary difficulty is that many large observation clinical studies do not follow all patients who are called negative by the diagnostic test. Some of these patients will truly be diseased, and the number of false negative patients and the true number of diseased patients in those studies are unknown. Often these observational studies only report rates of detection and recall. From these studies we cannot directly calculate some FOMs, like TPFs or FPFs, or their differences. However we can calculate or approximate the relative values of these statistics between study arms when the arms use the same patient population (39,40). Therefore, we quantify effects in terms of ratios of FOMs of a new modality to a standard modality.

MATERIALS AND METHODS

The TPF is the fraction of all diseased patients who are classified as positive by the diagnostic. Likewise, the FPF is the fraction of all nondiseased patients who are classified as positive by the diagnostic. Specificity or the true negative fraction (TNF) is simply the complement of the FPF, TNF = 1 − FPF, so an increase in FPF requires a decrease in specificity. Therefore, any inferences we make regarding the reproducibility of changes in FPF also apply to specificity. TPF and FPF are easily computed, so they are frequently reported in diagnostic imaging studies.

Most diagnostics do not just have positive or negative outputs. Diagnostics measure a concentration like a bio-marker, or an X-ray density like a scanner, or a reader's confidence of disease. Typically, if that measured value for a patient is greater than some threshold, we may consider that a positive diagnosis. Values below the threshold are interpreted as negative. However, that threshold can be changed, and as we change that threshold TPF and FPF also change. How TPF and FPF vary with respect to each other as the decision threshold changes is the ROC curve (20), an example of which is shown in Figure 1.

The AUC is frequently used as a summary measure of diagnostic accuracy. The AUC can be interpreted in many ways. It is the average sensitivity (TPF) over all specificities or the average specificity (TNF) over all sensitivities. It is also the probability that a diagnostic will correctly identify which of two randomly selected patients is diseased (7). AUC is frequently used as a measure of overall diagnostic accuracy.

"Expected utility" (EU) (41) is a measure of the trade-off between the TPF and the FPF at a clinical decision threshold. Specifically, it is defined as

$$EU = TPF_c - \beta \times FPF_c \quad (1)$$

where TPF_c and FPF_c are the TPF and FPF on a reader's ROC curve where the slope of the curve is β. Based on the clinical practice of mammography, Abbey et al. (26) determined