Estimating Screening-Mammography Receiver Operating Characteristic (ROC) Curves from Stratified Random Samples of Screening Mammograms:

A Simulation Study

Richard M. Zur, PhD[1], Lorenzo L. Pesce, PhD[2], Yulei Jiang, PhD

Rationale and Objectives: To evaluate stratified random sampling (SRS) of screening mammograms by (1) Breast Imaging Reporting and Data System (BI-RADS) assessment categories, and (2) the presence of breast cancer in mammograms, for estimation of screening-mammography receiver operating characteristic (ROC) curves in retrospective observer studies.

Materials and Methods: We compared observer study case sets constructed by (1) random sampling (RS); (2) SRS with proportional allocation (SRS-P) with BI-RADS 1 and 2 noncancer cases accounting for 90.6% of all noncancer cases; (3) SRS with disproportional allocation (SRS-D) with BI-RADS 1 and 2 noncancer cases accounting for 10%–80%; and (4) SRS-D and multiple imputation (SRS-D + MI) with missing BI-RADS 1 and 2 noncancer cases imputed to recover the 90.6% proportion. Monte Carlo simulated case sets were drawn from a large case population modeled after published Digital Mammography Imaging Screening Trial data. We compared the bias, root-mean-square error, and coverage of 95% confidence intervals of area under the ROC curve (AUC) estimates from the sampling methods (200–2000 cases, of which 25% were cancer cases) versus from the large case population.

Results: AUC estimates were unbiased from RS, SRS-P, and SRS-D + MI, but biased from SRS-D. AUC estimates from SRS-P and SRS-D + MI had 10% smaller root-mean-square error than RS.

Conclusions: Both SRS-P and SRS-D + MI can be used to obtain unbiased and 10% more efficient estimate of screening-mammography ROC curves.

Key Words: Screening mammography; simulation study; stratified random sampling; observer studies; ROC analysis.

Screening mammography for early detection of breast cancer leads to reduction in breast cancer mortality (1–3). Radiologists' interpretation of screening mammograms is paramount to the effectiveness of breast cancer screening. Receiver operating characteristic (ROC) analysis, which summarizes inherent trade-offs between sensitivity and specificity as the decision threshold is made more or less stringent, is an established method for the assessment of diagnostic performance. However, reliable estimation of ROC curves requires both the diagnostic "truth" (ie, whether breast cancer is present in the mammogram) and an ordinal response for every patient from the radiologist of an estimated "likelihood of malignancy." Therefore, ROC curves are usually estimated only in retrospective observer-performance studies (hereafter simply observer studies), in which readers provide likelihood of malignancy responses to cases of which diagnostic truth has been independently verified.

It is impractical to estimate screening-mammography ROC curves in observer studies by simple random sampling (RS) of clinical cases because of the low prevalence of breast cancer, approximately five per 1000 screening mammograms (4). This low cancer prevalence implies large uncertainty in the ROC curve estimate even if the total number of cases is large (5,6). Furthermore, 90% or more of all cases will be interpreted as either Breast Imaging Reporting and Data System (BI-RADS) assessment category 1 (negative) or 2

Acad Radiol 2015; 22:580–590

[1] Present address: The Hospital for Sick Children, 555 University Avenue, Toronto, ON, M5G 1X8, Canada.
[2] Present address: Computation Institute, Searle Chemistry Laboratory, 5735 South Ellis Avenue, Chicago, IL 60637.

Department of Radiology, The University of Chicago, 5841 South Maryland Avenue, MC2026, Chicago, IL, 60637 (R.M.Z., L.L.P., Y.J.). Received April 21, 2014; accepted December 9, 2014. This work was supported in part by the National Cancer Institute (NCI) of the National Institutes of Health (NIH) through grant CA092361. **Address correspondence to**: R.M.Z. e-mail: richard.zur@sickkids.ca

http://dx.doi.org/10.1016/j.acra.2014.12.011

TABLE 1. Observer Study Case Set Sampling Methods

Method	Abbreviation	Explanation
Random sampling	RS*	Random sample from large clinical case population
Stratified random sampling with proportional allocation	SRS-P*	Random sample within each stratum of (1) noncancer cases assigned BI-RADS 1 clinically, (2) noncancer cases assigned BI-RADS 2 clinically, (3) noncancer cases assigned BI-RADS 0 clinically, and (4) cancer cases, with fixed number-of-case ratios between noncancer-case strata identical to those of large clinical case population
Stratified random sampling with disproportional allocation	SRS-D	Random sample within each stratum as in SRS-P, with fixed number-of-case ratios between noncancer-case strata different from those of large clinical case population
Stratified random sampling with disproportional allocation with multiple imputation	SRS-D + MI	Random sample within each stratum as in SRS-P, with fixed number-of-case ratios between noncancer-case strata different from those of large clinical case population. After multiple imputation, the number-of-case ratios between noncancer-case strata becomes identical to those of large clinical case population

The proportion of cancer cases was fixed at 25% for all methods (before multiple imputation).

*The sampled case sets were similar between RS and SRS-P except that the number-of-case ratios between noncancer-case strata were exact (not subject to statistical sampling variation) with SRS-P but approximate (subject to statistical sampling variation) with RS.

(benign finding), leading to repetitive and uninteresting studies for the observers (7). Investigators often increase the prevalence of cancer cases in observer studies by including fewer noncancer cases then seen in clinical practice. This approach can greatly alleviate the difficulty caused by low cancer prevalence and increase the efficiency of the observer study by decreasing the uncertainty of ROC curve estimates without increasing the number of cases in the study.

In a Monte Carlo simulation study, we compared methods that use stratified random sampling (SRS) for the construction of observer study case sets based on (1) BI-RADS assessment category assigned clinically and (2) cancer versus noncancer truth status, including one method that uses multiple imputation (MI) for correction of bias caused by SRS (8). Our study shows that two of these methods can produce unbiased and more efficient estimates of screening-mammography ROC curves.

MATERIALS AND METHODS

We begin with a large population of clinical mammogram cases for which we wish to estimate the screening-mammography ROC curve. For example, in this study, we used approximately the 49,500 cases of the Digital Mammography Imaging Screening Trial (DMIST) (9). From this large population of clinical cases we build a smaller observer study case set. The goal is to construct the observer study case set to be as small as possible for efficient (low uncertainty) and unbiased estimation of screening-mammography ROC curves.

Stratified Random Sampling

With SRS, a case population is divided into nonoverlapping groups (or strata), and sample cases are drawn randomly from within each stratum and at the same time maintaining fixed number-of-case ratios between the strata. Strata can be defined in many ways; in general, one defines strata in such a way that minimizes the within-stratum variance of some parameters of interest (eg, readers' likelihood of malignancy score) compared to corresponding between-strata variance to reduce the overall estimation uncertainty (10).

We define three strata for noncancer cases: (1) those that are assigned BI-RADS assessment category 1 in the clinical interpretation, (2) those assigned BI-RADS 2, and (3) those assigned BI-RAD 0; and we define a single stratum for cancer cases. We expect the observer to report low likelihood of malignancy scores in most BI-RADS 1 and 2 noncancer cases.

In this study, we fixed the proportion of cancer cases at 25% in observer study case sets and considered three different SRS methods for noncancer cases (Table 1). (1) SRS with proportional allocation (SRS-P): the number-of-case ratios between noncancer-case strata are identical to those of the large clinical case population. (2) SRS with disproportional allocation (SRS-D): the number-of-case ratios between noncancer-case strata are different from those of the clinical case population and the proportion of BI-RADS 0 cases is relatively greater. (3) SRS-D with multiple imputation (SRS-D + MI): MI (described subsequently) is used to recover the number-of-case ratios between noncancer-case strata of the clinical case population, thereby correcting the effect of SRS-D on the number of cases (and in that process also correcting bias on ROC estimate caused by SRS-D). We compared the three SRS methods against the reference standard of RS, which is expected to produce unbiased but inefficient estimates. Bias in ROC estimation is the expected difference (ie, the average difference in a large number of repeated experiments) between the true and estimated ROC performance. Unbiased ROC estimates agree, on average, in a large number of repeated experiments, with the true ROC performance, and inefficient ROC estimates are characterized by large estimation variance.