*Short Communication*

# The Importance of Cytologic Intrarater and Interrater Reproducibility: the Case of Ductal Lavage

**Kala Visvanathan,[1,2,4] Deborah Santor,[1] Syed Z. Ali,[4] In Soon Hong,[5] Nancy E. Davidson,[2,4] and Kathy J. Helzlsouer[1,3]**

[1]The Johns Hopkins Bloomberg School of Public Health; [2]Sidney Kimmel Comprehensive Cancer Center; and [3]Mercy Medical Center, Baltimore, Maryland; [4]Johns Hopkins School of Medicine; and [5]Howard University, Washington, District of Columbia

## Abstract

**The reproducibility of a test result is a critical component of its clinical utility. Little information is available concerning the intrarater reproducibility of cytologic assessments. This study evaluated the reproducibility of cytologic interpretation of epithelial cells obtained from ductal lavage (DL), a minimally invasive method used to obtain sample cells from breast tissue. Two cytospin slides were made for each duct sampled. Slides with <10 cells were considered inadequate to make a diagnosis; the remaining slides were classified into mildly atypical, markedly atypical, and malignant cells. Each pair of slides were classified by the more serious diagnosis. DL samples from 100 ducts were independently blind-reviewed by two experienced cytopathologists. All abnormal slides and a random sample of normal slides and slides identified as inadequate for diagnosis ($n$ = 43) were re-reviewed. The $\kappa$ for intrarater agreement was 0.59 $\pm$ 0.10 for cytopathologist 1 and 0.33 $\pm$ 0.08 for cytopathologist 2. The $\kappa$ for interrater agreement of slides from 100 ducts was 0.46 $\pm$ 0.07. The interrater agreement of the slides that were re-reviewed was $\kappa$ = 0.27 $\pm$ 0.09. Fair to moderate intrarater and interrater agreement of DL cytology was observed. Low intrarater and interrater cytologic consistency may compromise the interpretation of clinical studies of DL.** (Cancer Epidemiol Biomarkers Prev 2006;15(12):2553–6)

## Introduction

Ductal lavage (DL) is a minimally invasive procedure used to obtain breast ductal epithelial cells that could potentially be used for breast cancer risk assessment, early detection, and evaluation of intermediate markers in chemoprevention trials. After the insertion of a small microcatheter into a breast duct, and gradual infusion of a few milliliters of saline (1), the breast is massaged and the effluent and ductal epithelial cells that have been dislodged from the epithelial surface are collected for analysis.

In the initial study published by Dooley et al., a cytologic classification for DL was developed by experienced cytopathologists (1). Epithelial cells were classified into four distinct categories: benign, mildly atypical, markedly atypical, and malignant. Slides with <10 cells were considered to have insufficient cellular material to make a diagnosis (ICMD; ref. 1). This classification is a modification of established NIH consensus criteria (1996) used for cytologic evaluations of fine-needle aspirations (FNA) of palpable and nonpalpable breast lesions (2). FNAs were classified into benign, atypical/indeterminate, suspicious/probably malignant, malignant, and unsatisfactory using the consensus criteria.

Subsequent DL studies using the Dooley et al. cytologic classification have reported low sensitivity, specificity, and reliability—raising concerns about its feasibility as a diagnostic and/or risk assessment tool (3, 4). Two cross-sectional studies compared mastectomy specimens with known carcinoma to DL cytology from the same individuals. In the first study, two out of three pathologists reported markedly atypical, but no malignant cells in 14% (4 of 29) of specimens (3). In a second study that used a higher diagnostic threshold of ≥100 epithelial cells, similar results were observed. Markedly atypical and malignant cells were diagnosed in 5 out of 38 specimens giving a sensitivity of 13% (95% confidence intervals, 6-29%). In six cancer-free specimens, a specificity of 100% was reported (95% confidence intervals, 54-100%; ref. 4). The sensitivity increased to 43% (95% confidence intervals, 23-72%) and specificity decreased to 96% (95% confidence intervals, 86-100%) when the comparison with histologic diagnosis was restricted to the same duct rather than to the entire breast. In a separate study, the reproducibility of mild or markedly atypical cells was evaluated in the unaffected breast of 23 women after a median period of 8.3 months (2.3-14.3). Among the 78 ducts that were re-lavaged, the reproducibility of atypia was only 19%. In those individuals who produced nipple aspirate, the reproducibility of atypia was higher at 55% (5).

A potentially contributing and modifiable factor for the observed low sensitivity, specificity, and reliability of DL cytology may be the lack of consistency in the cytologic interpretations within and between pathologists. Although three out of eight DL studies (1, 3, 4) assessed interrater cytologic reliability, none reported on intrarater reproducibility. In this study, we examine the intrarater reliability of DL cytologic assessments by two experienced cytopathologists, and interrater reliabilities twice.

## Materials and Methods

The study assessed both intrarater and interrater reliability using DL samples collected between May 2002 and April 2005

Received 7/12/06; revised 9/18/06; accepted 10/5/06.

**Grant support:** Avon Foundation, The National Institute of Health (5U01CA/ES62988), and The Maryland Cigarette Restitution Fund. Dr. Visvanathan is a recipient of a Breast Specialized Programs of Research Excellence Career Development Award (NIH CA 88843), an American Society of Clinical Oncology Career Development Award (ASCO), and KO7 Preventive Oncology Academic Award (NCI CA11948).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

**Requests for reprints:** Kala Visvanathan, Department of Epidemiology, Johns Hopkins University, Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205. Phone: 410-614-1112; Fax: 410-614-2632. E-mail: kvisvana@jhsph.edu



doi:10.1158/1055-9965.EPI-06-0578

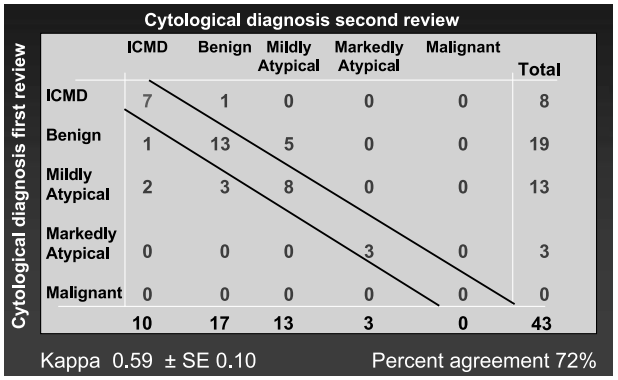


**Figure 1.** Intrarater agreement (pathologist 1).

from a clinical research study evaluating the reliability of nipple aspirate fluid, and DL 6 months apart in women at high risk for breast cancer. DL was only done on fluid-producing ducts. These women are currently being followed for a period of 2 years. Recruitment to the study was from a risk assessment and genetic counseling clinic as well as from a breast evaluation program at Johns Hopkins. Women were considered eligible for the study if they had (*a*) an estimated National Cancer Institute/Gail model lifetime risk of developing breast cancer of >20% or a 5 year risk of ≥1.66%; (*b*) a known BRCA1 or BRCA2 mutation; (*c*) a prior history of invasive or noninvasive breast cancer; (*d*) a prior history of breast hyperplasia with or without atypia; or (*e*) a family history of breast cancer defined as a first-degree relative or two second-degree relatives on the same side of the family with a history of breast cancer. All study participants had to be 18 years or older, sign an informed consent, have had a normal clinical breast exam, and, for women 40 and over, a nonsuspicious mammogram for malignancy in the 12 months prior to the date of enrollment. The Committee of Human Research at the Johns Hopkins Bloomberg School of Public Health approved the study.

Two cytospin slides fixed in 95% ethanol and stained with the Papanicolaou stain were made from each DL sample. All slides were blind-reviewed in pairs by two experienced clinical cytopathologists. Both cytopathologists had experience in reporting DL samples prior to the study, reviewed guidelines published by the Cytyc Health Corporation (http://www.medcyt.com/), and had extensive expertise in the interpretation of breast fine needle aspirates that use a similar cytologic classification to DL (1). The cellular component of each slide was initially evaluated for adequacy by conducting a semi-quantitative epithelial cell count. Slides with <10 epithelial cells were considered to be ICMD and were not examined further. The remaining slides were categorized into benign, mildly atypical, markedly atypical, and malignant. In cases in which there were two slides with 10 cells or more, the cytologic diagnosis was based on the most severe alteration identified. Slides that were considered abnormal (mildly atypical, markedly atypical, or malignant) by either cytopathologist ($n$ = 20), were re-reviewed by both. Along with the abnormal slides, a random sample of 50% of slides considered benign by both cytopathologists ($n$ = 14), and nine slides that were considered ICMD by both were also re-reviewed. The $\kappa$ statistic (6) was used to assess intrarater and interrater reliability.

## Results

Sixty-nine women enrolled in the study and 47 women returned for a second visit. The mean age of study participants was 46.6 years (SD, 7.9) and more than half were premenopausal (59%). Ninety-six percent of the study participants were Caucasians and the remaining 4% were African-Americans. Fifty-five percent had a Gail risk of ≥1.66, and 77% had a family history of breast cancer. Two of 14 women who had undergone clinical genetic testing for BRCA1 or BRCA2 mutations were found to be positive for a deleterious mutation. Eleven women had prior surgery for either *in situ* or invasive breast cancer. In five of these women, DL was only done on the unaffected breast. Six women had breast conservation and radiation treatment, none of these women produced nipple aspirate fluid, and therefore, DL was not done on the affected breast. Ten women reported a prior diagnosis of atypical hyperplasia, from which five underwent DL of the affected breast. None of the six women on tamoxifen (three for treatment of breast cancer and three for prevention) or the three women on raloxifene underwent DL.

Lavage fluid was successfully collected from 58 ducts in 29 women at visit 1 and 42 ducts in 18 women at visit 2, producing a total of 100 samples to review. The re-review of 43 slides (abnormal, $n$ = 20; benign, $n$ = 14; ICMD, $n$ = 9) by both cytopathologists produced an intrarater agreement of $\kappa$ = 0.59 ± 0.10 (SE) for cytopathologist 1, and $\kappa$ = 0.33 ± 0.08 (SE) for cytopathologist 2 (see Figs. 1 and 2). For each cytopathologist, the intrarater agreement on distinguishing ICMD versus no ICMD was 80% or higher. The $\kappa$ for cytologic interrater agreement on initial review ($n$ = 100) was 0.46 ± 0.07 and 0.27 ± 0.09 for the slides that were re-reviewed ($n$ = 43; see Figs. 3 and 4). To date, all but five women have been followed for the complete 2-year period and none of the women have developed breast cancer.

## Discussion

The adverse effect of either a false positive or the false reassurance from a negative test in the context of breast cancer risk assessment and/or early detection, underscores the necessity for a high level of intrarater and interrater reliability for procedures such as DL. In this study, fair to moderate interrater agreement for DL cytology was observed. Previous studies have reported only on interrater reliability for DL cytology, two in women with breast cancer and one in women at high risk for developing breast cancer (1, 3, 4). Two of the studies that used the same cytologic classification reported similar results (1, 3). In one study, 29 specimens were reviewed by three cytopathologists; the average weighted $\kappa$ of each pair of cytopathologists was 0.52 (3). A further review of 60 specific cellular characteristics in 431

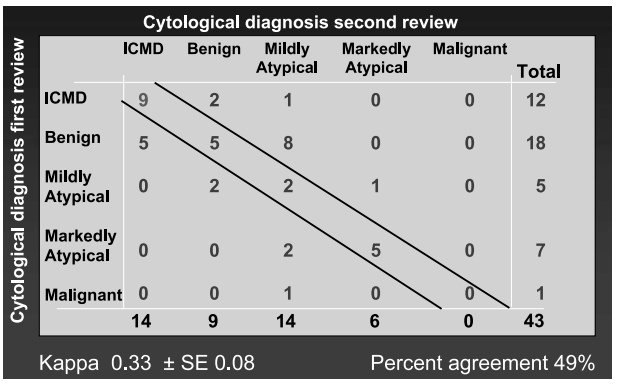


**Figure 2.** Intrarater agreement (pathologist 2).

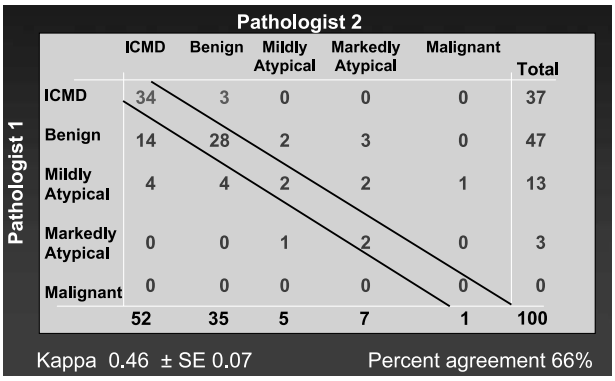


**Figure 3.** Interrater agreement (review 1).

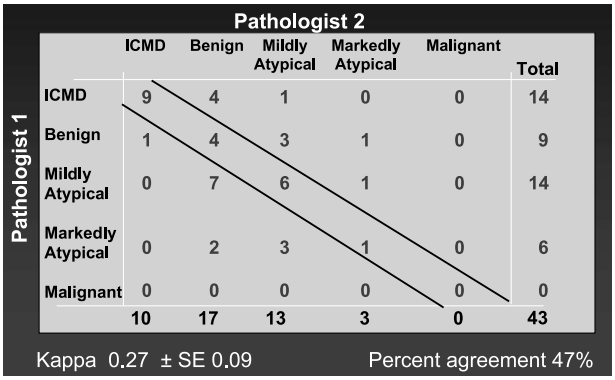


**Figure 4.** Interrater agreement (review 2).

DL specimens by the two cytopathologists involved in the initial study by Dooley et al. has also been published (7). This group had initially reported a concordance rate of 89% between cytopathologists. However, the degree of agreement on detailed review was only substantial ($\kappa$ >0.70) for 5 of the 60 characteristics: increase in nuclear size, absent or mild anisonucleosis, increased nucleoli size, mitosis, and necrotic debris. In a third study, the weighted $\kappa$ was higher at 0.70 (95% confidence intervals, 0.54-0.86) for 52 samples reviewed by two cytopathologists, but the cellular threshold considered adequate for making a cytologic diagnosis was 100 epithelial cells per slide, making a comparison difficult (4). In all three studies that assessed interrater reliability, one of the reviewers was an experienced cytopathologist involved in devising the initial classification (1, 3, 4).

The results of interrater agreement in three published studies of diagnostic FNA independent of information from a clinical examination or imaging studies reported similar results. In one study, 20 FNAs of clinically palpable lesions were reviewed by six pathologists. The degree of agreement among six pathologists was highest for malignant cells ($\kappa$ = 0.75) and lowest for atypical cells ($\kappa$ = 0.08; ref. 8). In a second study, 12 breast lesions ranging from nonproliferative to low-grade ductal carcinoma *in situ*, were reviewed by six pathologists and the weighted $\kappa$ was 0.35 (9). A third study involving the blind review of 41 false negative and 49 true negative FNA specimens by 10 cytopathologists resulted in $\kappa$ = 0.54 (range, 0.40-0.65; refs. 10). In all three studies, FNA was used to evaluate abnormal lesions which are typically associated with higher cellularity, compared with our study where we had mostly benign cytology, which has lower cellularity. It is likely that FNA also performs poorly on normal breast tissue, which has low cellularity. Therefore, it is critical to also examine the reliability of random periareolar FNA, which has been proposed as a means to evaluate breast tissue response to chemopreventive interventions (11).

Intrarater reliability is also a major component that will contribute to the overall accuracy of the assessment; however, this component has not been previously examined in DL. Using the search terms' reliability, intrarater reliability, or intraobserver variation of FNA of the breast in MEDLINE, we did not identify any published data on the intrarater reliability of FNA. In this study, despite experienced cytopathologists trained in FNA and DL, only moderate cytologic intrarater agreement was observed.

Further studies are needed to determine if intrarater and interrater reliability of DL cytology can be improved by more extensive formal training, changes to the classification, or increases in cellular yield (12, 13). Formal training in sample procurement and reading of slides has been shown to improve the sensitivity and specificity of FNA interpretation. This study, however, did not evaluate interrater and intrarater reliability. The cytopathologists evaluating DL cytology in the current study had both undergone formal training on FNA and DL cytology interpretation; a lower degree of interrater agreement was observed on second review of an enriched subset of slides that included all abnormal slides from the first review.

Our study highlights the importance of incorporating a comprehensive reliability assessment including both intrarater and interrater reliability measures into evaluations of procedures even for accepted clinical testing such as cytologic evaluation. Our results suggest that a lack of internal consistency of cytologic interpretation is a major problem to be dealt with before translating DL cytology to more widespread clinical applications. Potential approaches to improve reliability include additional training or modification of existing diagnostic criteria. Alternatively, if such measures are ineffective, early identification could help to assess the viability of continued evaluation of a procedure from both scientific and cost-effective perspectives. These results also indicate that studies evaluating biomarkers present in ductal fluid should validate their assays against the gold standard of histologic confirmation of the lesion by tissue examination rather than cytology.

## Acknowledgments

We thank Anita Mayfield for assistance with slide preparation and Ram Tenkas: for privinding helpful comments on the article.

## References

1. Dooley WC, Ljung BM, Veronesi U, et al. Ductal lavage for detection of cellular atypia in women at high risk for breast cancer. J Natl Cancer Inst 2001;93:1624–32.
2. The uniform approach to breast fine-needle aspiration biopsy. NIH consensus development conference. Am J Surg 1997;174:371–85.
3. Brogi E, Robson M, Panageas KS, Casadio C, Ljung BM, Montgomery L. Ductal lavage in patients undergoing mastectomy for mammary carcinoma: a correlative study. Cancer 2003;98:2170–6.
4. Khan SA, Wiley EL, Rodriguez N, et al. Ductal lavage findings in women with known breast cancer undergoing mastectomy. J Natl Cancer Inst 2004; 96:1510–7.
5. Johnson-Maddux A, Ashfaq R, Cler L, et al. Reproducibility of cytologic atypia in repeat nipple duct lavage. Cancer 2005;103:1129–36.
6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.

7. Ljung BM, Chew KL, Moore DH II, King EB. Cytology of ductal lavage fluid of the breast. Diagn Cytopathol 2004;30:143–50.
8. Gornstein B, Jacobs T, Bedard Y, et al. Interobserver agreement of a probabilistic approach to reporting breast fine-needle aspirations on ThinPrep. Diagn Cytopathol 2004;30:389–95.
9. Sidawy MK, Stoler MH, Frable WJ, et al. Interobserver variability in the classification of proliferative breast lesions by fine-needle aspiration: results of the Papanicolaou Society of Cytopathology Study. Diagn Cytopathol 1998; 18:150–65.
10. Bulgaresi P, Cariaggi MP, Bonardi L, et al. Analysis of morphologic patterns of fine-needle aspiration of the breast to reduce false-negative results in breast cytology. Cancer 2005;105:152–7.
11. Fabian CJ, Kimler BF, Mayo MS, Khan SA. Breast-tissue sampling for risk assessment and prevention. Endocr Relat Cancer 2005;12:185–213.
12. Ljung BM, Drejet A, Chiampi N, et al. Diagnostic accuracy of fine-needle aspiration biopsy is determined by physician training in sampling technique. Cancer 2001;93:263–8.
13. Brogi E, Miller MJ, Casadio C, Lyung BM, Montgomery L. Paired ductal lavage and fine-needle aspiration specimens from patients with breast carcinoma. Diagn Cytopathol 2005;33:370–5.

# The Importance of Cytologic Intrarater and Interrater Reproducibility: the Case of Ductal Lavage

Kala Visvanathan, Deborah Santor, Syed Z. Ali, et al.

*Cancer Epidemiol Biomarkers Prev* 2006;15:2553-2556.