Published in final edited form as:
Arch Pathol Lab Med. 2011 December ; 135(12): 1557–1560. doi:10.5858/arpa.2010-0739-OA.

Prevalence of Abnormalities Influences Cytologists' Error Rates in Screening for Cervical Cancer

Dr. Karla K. Evans, PhD, **Dr. Rosemary H. Tambouret, MD**, **Andrew Evered Mr., BSc**, **Dr. David C. Wilbur, MD**, and **Dr. Jeremy M. Wolfe, PhD**
Department of Ophthalmology, Visual Attention Laboratory, Brigham and Women's Hospital, Harvard Medical School, Cambridge, Massachusetts (Drs Evans and Wolfe); the Department of Pathology, Massachusetts General Hospital, Boston (Drs Tambouret and Wilbur); and the Department of Pathology, Cervical Screening, University of Wales Institute Cardiff, Wales (Mr Evered)

Abstract

Context—Medical screening tasks are often difficult, visual searches with low target prevalence (low rates of disease). Under laboratory conditions, when targets are rare, nonexpert searchers show decreases in false-positive results and increases in false-negative results compared with results when targets are common. This *prevalence effect* is not due to vigilance failures or target unfamiliarity.

Objective—To determine whether prevalence effects could be a source of elevated false-negative errors in medical experts.

Design—We studied 2 groups of cytologists involved in cervical cancer screening (Boston, Massachusetts, and South Wales, UK). Cytologists evaluated photomicrographs of cells at low (2% or 5%) or higher (50%) rates of abnormality prevalence. Two versions of the experiment were performed. The Boston, Massachusetts, group made decisions of normal or abnormal findings using a 4-point rating scale. Additionally, the group from South Wales localized apparent abnormalities.

Results—In both groups, there is evidence for prevalence effects. False-negative errors were 17% (higher prevalence), rising to 30% (low prevalence) in the Boston, Massachusetts, group. The error rate was 27% (higher prevalence), rising to 42% (low prevalence) in the South Wales group. (Comparisons between the 2 groups are not meaningful because the stimulus sets were different.)

Conclusions—These results provide the first evidence, to our knowledge, that experts are not immune to the effects of prevalence even with stimuli from their domain of expertise. Prevalence is a factor to consider in screening for disease by human observers and has significant implications for cytology-based cervical cancer screening in the post–human papillomavirus vaccine era, when prevalence rates of high-grade lesions in the population are expected to decline.

Routine medical screening can save lives.[1–3] A characteristic of most screening tasks is that cases of the disease are rare. In cervical cancer screening by cytology, rates of finding any abnormal cells in a well-screened population are usually around 5% or less in the United States, and the prevalence of cells of the most significant high-grade disease is less than

Reprints: Karla K. Evans, PhD, Department of Ophthalmology, Visual Attention Laboratory, Brigham and Women's Hospital, Harvard Medical School, 64 Sidney St, Ste 170, Cambridge, MA 02139, kevans@search.bwh.harvard.edu.

The authors have no relevant financial interest in the products or companies described in this article.

Presented as a poster at the 58th Annual Scientific Meeting of the American Society of Cytopathology, Boston, Massachusetts, November 13–15, 2010.

1%.[4] The National Health and Social Care Information center in the United Kingdom reported in 2009 that, in well-screened populations, typically 5% to 6% of adequate samples contain mild abnormalities and 1% to 2% will have a severe abnormality. Under laboratory conditions, when targets are rare, non-expert searchers show increases in false-negative and decreases in false-positive errors compared with conditions where targets are common. Thus, using nonmedical stimuli, observers are more likely to miss the target if it is rare than if it is common. Furthermore, the observed prevalence effect is not due to vigilance failures or unfamiliarity with the target.[5] If this prevalence effect from the laboratory also applies to experts, this could be an important contributor to medical error.

Cervical screening by cytologic examination of cells is one such example of medical screening where low target prevalence is a characteristic of the visual search task. Cytologists routinely screen samples daily in which the prevalence rate of the most significant pathology is usually less than 1%.[4] Cytology audits suggest that low numbers of abnormal cells in individual cervical cytology slides are associated with false-negative reporting.[6,7] In the era of human papillomavirus vaccination, the effects of declining disease prevalence on the performance of cytologic screening should be considered because the introduction of the vaccine is expected to eliminate significant numbers of high-grade lesions from the population.[8] Here, we report 2 independent studies that test whether low prevalence of abnormal findings in cytologic screening can be a factor in increased error rates in clinical settings. We compared the cytologists' performances in 2 conditions: (1) *low prevalence*, where abnormal cases are rare in a given sample of cases as would be the case in standard clinical screening, and (2) *high prevalence*, where the abnormalities were present in 50% of the cases.

MATERIALS AND METHODS

Participants

Participants recruited for this study were 2 groups of cervical cytologists, 10 from Boston, Massachusetts, and 12 from Cardiff, South Wales. Each participant reported 20/25 vision or better and no history of eye or muscle disorders. All observers gave informed consent, as approved by the appropriate institutional human subjects review boards, and were compensated for their time. The group of cytologists from South Wales consisted of cytology screeners and biomedical scientists who met the quality standards of Cervical Screening Wales and the National Health Service Cancer Screening Programs of the United Kingdom. All participants possessed the National Health Service Cancer Screening Programs Certificate in Cervical Cytology or equivalent. Ten recruits were women, with ages between 31 and 55 years, and 2 recruits were men, aged 31 and 47 years. All recruits had, on average, 10 years cervical cytology screening experience. The group of cytologists from Boston, Massachusetts, consisted of 10 cytotechnologists who were certified by the American Society for Clinical Pathologists Board of Registry. Eight recruits were women, with ages between 27 and 68 years, and 2 recruits were men, aged 41 and 47 years. All recruits were actively engaged in the daily practice of laboratory cervical cytology screening and had at least 3 years of experience, with an average of 18 years.

Stimuli and Apparatus

All images used in the study were glass-slide samples of cervical cytology prepared using SurePath liquid-based technology (Becton-Dickinson, Burlington, North Carolina), stained with the Papanicolaou stain.

The study conducted in South Wales used 90 images of abnormal cytology and 1000 cytologic images without abnormality. The cytology images without abnormality were shown twice, randomly throughout the experiment, but the second time they were rotated by 180°. Slides with cytologic abnormalities were a mix of low-grade and high-grade squamous intraepithelial lesions, independently prevalidated by 3 expert cytologists. Glandular cell abnormalities were not included. The glass-slide samples were examined and photographed at ×100 magnification (×10 ocular and ×10 objective) using an Olympus BX51 optical microscope (Olympus, Tokyo, Japan) equipped with a ColorView II digital camera (Soft Imaging System Ltd, Helperby, North Yorkshire, England). All images were acquired using analySIS software (Soft Imaging System) at a resolution of 2010 × 51 544 pixels. Participants examined the images on a 19-inch, liquid-crystal color display screen set at a resolution of 1024 × 768 pixels.

For the Boston, Massachusetts, study, at least 2 cytopathologists concurred on the interpretation of all images: 1950 slides were interpreted as normal, 100 as low-grade squamous intraepithelial lesions, and 50 as high-grade squamous intraepithelial lesions. Images were acquired using an Olympus BX40CY microscope and photographed at ×200 magnification (×10 ocular and ×20 objective) using a Spot Insight Color digital camera v3.4 (Diagnostic Instruments, Inc, Sterling Heights, Michigan). All 150 images of abnormal cytology (a mixture of low-grade squamous intraepithelial lesion and high-grade squamous intraepithelial lesion) were used in the study. Sixty images of cytology without abnormality were shown twice during the experiment, but the second time they were rotated 180°, and 1890 slides were shown once. The participants viewed the images on a 21-inch cathode ray tube color monitor set at a resolution of 1024 × 768 pixels. Both of the studies were constructed and presented using the matrix laboratory (MATLAB, MathWorks, Natick, Massachusetts) program and Psychophysics Toolbox (http://psychtoolbox.org).[9,10]

Procedure

Two versions of the experiment were performed in which experts evaluated images of cells at either low (2% in South Wales or 5% in United States) or higher (50%) prevalence of abnormality. We use the 2 low-prevalence conditions because they mimic the prevalence in clinical screening. High prevalence of 50% is typical of many laboratory experiments in this area. Moreover, training tends to occur at high prevalence. One could examine the full range of prevalence values,[11] but that would require, in this case, much more time from each observer without producing much additional information. Images were displayed on the computer monitor, one at a time. The observers pressed the number 1 (definitely normal, no need for further review) to 4 (definitely abnormal, needs review) to indicate how certain they were about presence of an abnormality and the need to send the image for further review. Once a cytologist responded, the answer was confirmed by pressing the return button, or the participant had a chance to alter the rating. After confirmation, the program advanced to the next image. In South Wales, observers also recorded the perceived abnormalities by clicking on the locations in the image before proceeding to the next image. Observers were not given any image-by-image feedback. In addition to recording the ratings, time to response was recorded, and, in South Wales, the localization responses were also recorded.

Data collection took up to 8 hours per observer, during which, each observer saw and evaluated 2100 sample images. The trials were divided into 21 experimental blocks with 100 images. Twenty blocks had low target prevalence: 2% (40 out of 2000 images) in South Wales, and 5% (100 out of 2000 images) in the United States. Individual blocks varied in the number of targets present, so observers could not guess that their findings for a block were complete after finding 2 targets. In the one higher-prevalence block, observers saw 50 target-present and 50 target-absent images. This block was positioned exactly halfway through the experiment because we wanted to see if we could observe any effects of introducing a block of high prevalence on subsequent low prevalence blocks. We positioned the high-prevalence block in the middle of the experiment so we would have the same number of low-prevalence blocks before and after the high-prevalence block. Observers were allowed and encouraged to take short breaks.

Data Analysis

The main outcome measures or dependent variables were the miss (false-negative) and false-alarm (false-positive) error rates as a function of target prevalence. These error rates can be combined into the signal-detection measures of d' (the discriminability index), which indexes the ability of observers to tell the difference between target-present and target-absent images and c (criterion or bias), which gives a measure of the tendency of observers to respond positively or negatively. We report performance in d' values for 2 reasons. First, d' is theoretically independent of an observer's bias to respond "yes" or "no." Second, it is normally distributed, unlike accuracy, which makes it more suitable for standard parametric statistics. We report the criterion of the observers because it allows us to see the cutoff value determined by the observer trying to detect the signal at which the observer is ready to call something a reliable signal.

RESULTS

As can be seen in the Figure, a prevalence effect occurs when experts search for images in their domain of expertise. Figure, A, shows the US results. Figure, B, shows the results from South Wales. Of greatest practical interest, the miss-error rates in both experiments (blue line, Figure) were strongly affected by prevalence. In the US study, the average miss rate was 17% at high prevalence and 28% at low prevalence. In South Wales, the miss rate was 27% at high prevalence and 42% at low prevalence. The effect of prevalence is significant in both studies (Boston 2-tailed $t(9) = 8.12$, $P < .001$; South Wales 2-tailed $t(11) = 5.20$, $P < .001$). Note that performance between sites cannot be meaningfully compared because the stimuli and some of the methods were different.

False alarms (red) vary in the opposite direction, but false alarms are rare in these experiments. The effect of prevalence on criterion is also significant. The movement of miss errors in one direction and false alarms in the other is the standard sign of a criterion shift, so it is not surprising that the criterion measure, c, is significantly influenced by prevalence (Boston 2-tailed $t(9) = 5.13$, $P < .001$; South Wales 2-tailed $t(11) = 3.56$, $P = .004$). The measure of the ability of the observer to distinguish positive from negative images (D') does not change with prevalence. This is obvious in the US data (green line; Figure A) (2-tailed $t(9) = 0.52$, $P =.62$). Measures of d' have bigger variations in the scores obtained from South Wales but, again, the effect of prevalence is not significant (2-tailed $t(11) = 2.08$, $P = .06$).

In the laboratory setting, a block of high-prevalence trials with feedback shifts the criterion for subsequent low-prevalence trials in a manner that reduces miss errors.[5] Even without feedback, there is a hint of such an effect present in the current experiments. The 200 low-prevalence trials *after* the high-prevalence block have a lower miss-error rate than the 200 trials *before* (22% after versus 30% before the high-prevalence block), but that difference is not significant (2-tailed $t(9) = 1.24$, $P = .24$).

CONCLUSIONS

This study shows that target prevalence can influence the behavior of experts viewing stimuli in their domain of expertise. In this case, the ability of trained, practicing cytologists to identify the abnormal cells that must necessarily be found for optimal cervical cancer-screening programs. Specifically, false-negative rates are higher, and false-positive rates are lower at low prevalence. Prevalence of abnormal cells is already low in a cervical cancer-screening scenario, approximately 5% in most routine clinical populations, but much lower at about 1% for the most important high-grade lesions, those most likely to progress to cervical cancer if missed in the screening program. The current study, although not exactly mimicking laboratory microscopic screening, tests the hypothesis that increasing or decreasing the prevalence rate will affect the overall accuracy of cervical cancer screening in ways similar to those already noted in “nonexpert” task scenarios.[11] This study does not show that prevalence is a cause of medical error, but as long as there are humans involved in screening for disease, it is important to understand how their behavioral responses might influence the outcome of these tasks.

With the widening acceptance of vaccines targeting human papillomavirus types most common in high-grade cervical lesions, it seems likely those lesions will become less prevalent.[12] Early data from Australia, where high penetrance of the vaccine has already occurred in school-aged girls, shows that, in the vaccinated population, decreases in high-grade squamous intraepithelial lesion are already occurring, when compared with nonvaccinated cohorts.[13] Given that the present data support a prevalence/sensitivity relationship, methods of increasing the prevalence of abnormal cells to cytology screeners will be essential if the current efficacy of the program is to be maintained. Methods might include “seeding” of routine case populations with known abnormal slides, the use of automated screening devices with presentation of only high-probability fields of view to observers, or using molecular markers of high-grade disease which “stand out” from the slide background to guide screeners to cells of interest.[8] Alternatively, screening policymakers may wish to consider ways to screen for early cervical disease that do not rely on the visual reading of cell preparations. Human papillomavirus testing is one such method, and several trials have shown promising results in this respect.

Cervical cancer screening by cytology is, to our knowledge, the most successful cancer-prevention program devised. Cognizance of changes that may take place in its operational aspects with the introduction of vaccines is important so that modifications can take place to mitigate the prevalence effects documented here.

Acknowledgments

This work was supported by NIH EY017001 and ONR MURI N000141010278 to J.M.W.; K.K.E. was supported by NIH/NEI 1F32EY019819-01; and A.E. was supported by Research Capacity Building Collaboration Wales and British Society for Clinical Cytology and the National Association of Cytologists.

References

1. Brawley OW, Kramer BS. Cancer screening in theory and in practice. J Clin Oncol. 2005; 23(2): 293–300. [PubMed: 15637392]
2. Duffy SW, Tabar L, Olsen AH, et al. Absolute numbers of lives saved and overdiagnosis in breast cancer screening, from a randomized trial and from the Breast Screening Programme in England [published correction appears in *J Med Screen*2010;17(2):106]. J Med Screen. 2010; 17(1):25–30. [PubMed: 20356942]
3. Peto J, Gilham C, Fletcher O, Matthews FE. The cervical cancer epidemic that screening has prevented in the UK. Lancet. 2004; 364(9430):249–256. [PubMed: 15262102]
4. Benard VB, Eheman CR, Lawson HW, et al. Cervical screening in the National Breast and Cervical Cancer Early Detection Program, 1995–2001. Obstet Gynecol. 2004; 103(3):564–571. [PubMed: 14990422]
5. Wolfe JM, Horowitz TS, Van Wert MJ, Kenner NM. Low target prevalence is a stubborn source of errors in visual search tasks. J Exp Psychol Gen. 2007; 136(4):623–638. [PubMed: 17999575]
6. Mitchell H, Medley G. Differences between Papanicolaou smears with correct and incorrect diagnoses. Cytopathology. 1995; 6(6):368–375. [PubMed: 8770538]

7. O'Sullivan JP, A'Hern RP, Chapman PA, et al. A case control study of true positive versus false-negative cervical smears in women with cervical intraepithelial neoplasia (CIN) III. Cytopathology. 1998; 9(3):155–161. [PubMed: 9638376]

8. Schiffman M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S. Human papillomavirus and cervical cancer. Lancet. 2007; 370(9590):890–907. [PubMed: 17826171]

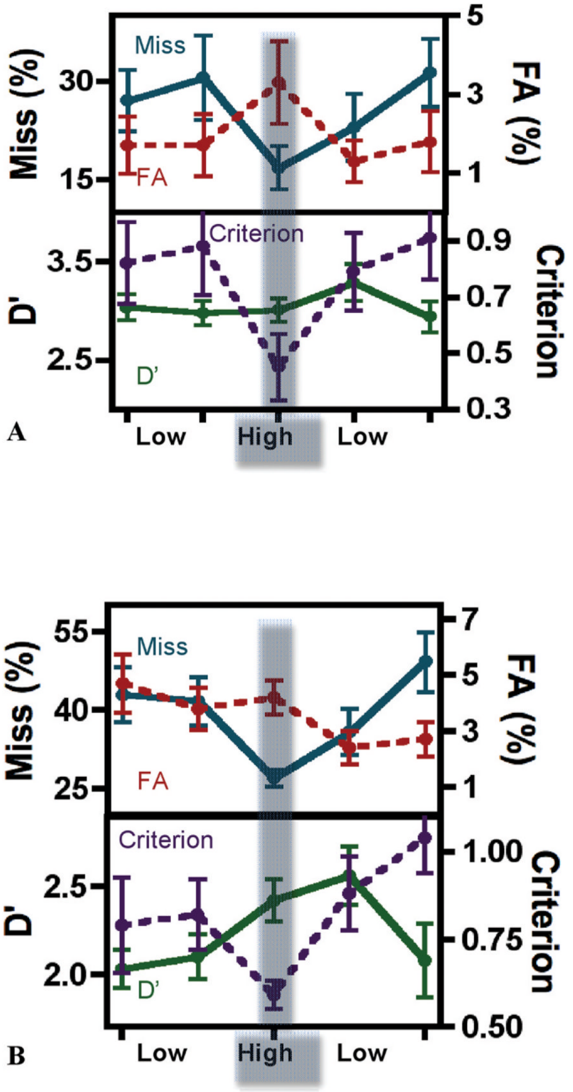
9. Brainard DH. The psychophysics toolbox. Spat Vis. 1997; 10(4):433–436. [PubMed: 9176952]

10. Pelli DG. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis. 1997; 10(4):437–442. [PubMed: 9176953]

11. Wolfe JM, Van Wert MJ. Varying target prevalence reveals two dissociable decision criteria in visual search. Curr Biol. 2010; 20(2):121–124. [PubMed: 20079642]

12. Franco EL, Cuzick J, Hildesheim A, de Sanjosé S. Chapter 20: issues in planning cervical cancer screening in the era of HPV vaccination. Vaccine. 2006; 24(suppl 3):S171–S177.

13. Brotherton, J. Surveillance of HPV vaccine impact. Paper presented at: 40th Annual Scientific and Business Meeting of the Australian Society of Cytology; October 17, 2010; Melbourne, Australia. North Ryde, New South Wales, Australia: Australian Society of Cytology Inc; 2010.



1. .

A and B, Top graphs are plots of average *z*-score false-negative results (blue solid line) and average *z*-score false-positive results (red dotted line) for low-target and high-target prevalence blocks. The bottom graphs are plots of average *D*′ (green solid line) and average *c* (criterion, purple dotted line) for blocks of low-target and high-target prevalence. A, Graphs depict data from the Boston, Massachusetts, study. B, Graphs depict data from the Cardiff, South Wales study. The 5 points on the graph represent data averaged over the 5 groups of blocks. The 5 groups of blocks are composed of 4 low-prevalence condition groups (each group is composed of 5 blocks of trials) and one high-prevalence block of trials. The error bars on the graphs are standard errors of the mean. Abbreviations: FA, false alarm or false-positive; Miss, false-negative, missed targets.