Use of Screening Phase Data to Evaluate Observer Variation of Sputum Cytodiagnosis as an Outcome Measure in a Chemoprevention Trial

George P. Browman, Andrew Arnold, Mark N. Levine, Terry D’Souza, Linda M. Turner-Smith, Romy Cayco, Brenda Johnstone, and Rhonda Birse

Ontario Cancer Treatment and Research Foundation, Hamilton Regional Centre, St. Joseph’s and Hamilton Civic Hospitals, and McMaster University, Hamilton, Ontario L8V 1C3, Canada

ABSTRACT

Sputum samples obtained during the screening phase of a chemoprevention trial in heavy smokers were evaluated independently by trained cytotechnologists and classified for degree of cellular atypia according to the method of Saccomanno et al. (G. Saccomanno et al., Cancer (Phila.), 33: 256–270, 1974). The level of agreement within and between Observers A and B was calculated as the percentage of agreement and, in addition, a statistic was used (κ) to correct for chance-expected agreement. Between observer agreement on 300 specimens from 130 subjects was 68% (204 of 300) (κ = 0.58). Of the 96 disagreements, only 17 were of more than one category in the six-category classification. Within observer agreement for both Observers A and B was evaluated on a subset of 60 specimens from 49 subjects examined on two separate occasions by each observer. The percentage of within observer agreement was 80% for Observer A (κ = 0.73) and 62% for Observer B (κ = 0.49) (P < 0.04). Altogether, 71% (25 of 35) of within observer discordant readings were confined to only one category. These data, obtained from the screening phase of the study, will allow us to refine the outcome measure for the trial without introducing bias that could result from knowledge of the actual study results.

The failure of conventional cancer treatments to impact significantly on overall cancer mortality has led to an emphasis on prevention. If premalignant changes can be reliably detected at a time when the malignant process is reversible, then it may be feasible to intervene to prevent the occurrence of some cancers. The appropriate selection of subjects for such intervention trials and the subsequent demonstration of the efficacy of chemoprevention strategies are therefore crucial.

INTRODUCTION

Based on the reported results of pilot studies (1, 2), we initiated a triple-blind randomized trial of the vitamin A analogue, etretinate, compared with placebo therapy in cigarette smokers with abnormal, premalignant sputum cytology (3). Since sputum cytology is being used as the primary outcome measure for this chemoprevention trial, it is important that its reproducibility be evaluated under conditions that will prevail during the course of the study. This includes assessment of the performance of the diagnosticians responsible for reporting the data. The purpose of this study was to refine the outcome measure and reporting mechanism for the trial by determining how well potential candidates for reviewing sputum cytology performed in evaluating specimens. Specimens obtained during the screening phase of the study were used to evaluate performance. We were chiefly concerned at this phase of the study with the reproducibility both within and between observers for classifying sputum cytology.

MATERIALS AND METHODS

Trial Design. The clinical study is a triple blind randomized trial in which subjects with at least a 15 pack-yr smoking history, who are active smokers, are determined to be eligible during a “run-in” period by screening for cellular atypia in the sputum (3). Screening for cellular atypia includes a requirement for submitting 3-day sputum specimens weekly on three separate occasions. The run-in phase is used also to identify subjects who are likely to be compliant.

Eligible subjects are randomized to receive either p.o. etretinate (25 mg) daily or an identical placebo, both for 6 mo. Follow-up sputum samples are collected monthly and read by observers unaware of treatment assignment.

Outcome Measure. Treatment groups will be compared with respect to the proportion of subjects showing reversal of sputum cellular atypia, according to the following definitions for treatment “response”: (a) for mild sputum atypia, a change to at least “metaplasia without atypia” (one-category change); (b) for “moderate atypia,” at least a two-category improvement to “no atypia”; and (c) for “marked atypia,” at least a two-category improvement to “mild atypia.” These definitions were chosen as being credible clinically and unlikely to be associated with misinterpretation of the cytological material.

Sample Preparation. The sputum samples evaluated in this reproducibility study were obtained from subjects during the screening phase of the trial prior to randomization. Each subject submitted a deep cough specimen daily for 3 consecutive days for 3 wk. Each 3-day collection was pooled and constituted a single specimen. Specimens were fixed and prepared according to the method of Saccomanno et al. (4) and stained using the Papanicolaou technique (5). Six slides were prepared per specimen.

Sample Reading. The sputum samples were prepared on glass slides which were coded and ordered randomly. During the screening phase of the study, two cytotechnologists (Observers A and B) prospectively and independently recorded the diagnosis using all six slides on each specimen submitted; between observer agreement was determined from these data. A subset of the original group of specimens was then selected, recoded, and resubmitted to the observers for independent classification on two further occasions at least 2 wk apart; these data were used to calculate within observer agreement. Specimens were graded according to the method of Saccomanno et al. (6) using the following categories: unsatisfactory specimen (either inflammatory or not a deep cough); satisfactory specimen without metaplasia; normal metaplasia without atypia; metaplasia with mild atypia; metaplasia with moderate atypia; metaplasia with marked atypia; and malignant cells present. Unsatisfactory specimens were those that contained inflammatory elements sufficient to obscure the cells to be examined, contained epithelial cells thought to be affected by an inflammatory process, or contained no alveolar macrophages. Specimens in which there were disagreements noted between observers were submitted to a diagnostic cytopathologist for adjudication. All observers were appropriately trained and registered cytotechnologists.

Statistical Considerations. Agreement was determined in two ways. (a) The percentage of agreement was calculated as the proportion of the number of complete agreements observed among the total number of judgments multiplied by 100%. (b) The κ statistic was used to express the extent of agreement beyond that expected by chance (7). The formula for κ is

\[ \kappa = \frac{Po - Pc}{1 - Pc} \]

where Po is the observed proportion of agreement, and Pc is the proportion of agreement expected by chance. κ has a value of 0 if observed agreement equals chance-expected agreement, +1 if observed agreement is perfect, and ≤0 if agreement is less than chance. Given that κ takes a value of 0 when Po = Pc, our data were used to calculate Po - Pc.

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.

1 Supported by a grant from the Ontario Ministry of Health.
2 To whom requests for reprints should be addressed, at Hamilton Regional Cancer Centre, 711 Concession St., Hamilton, Ontario L8V 1C3, Canada.
agreement is perfect, and <0 if observed agreement is below that expected by chance. Within observer $\kappa$ scores were compared statistically using the Z test (two-sided) (8).

RESULTS

Between Observer Agreement. There were 300 screened samples selected for evaluation from 130 subjects that represented the first 175 and the last 125 specimens screened between July 1, 1986, and July 1, 1987. These two cohorts were chosen to search for possible changes in performance over time.

Table 1 shows the agreement matrix for the 6-category cytology classification system for all 300 specimens. Complete agreements are represented along the diagonal of this 6 x 6 table. There was complete agreement on 204 of 300 (68%) samples. When chance-expected agreement is accounted for, overall agreement is 58% ($\kappa = 0.58$). The percentage of agreement and $\kappa$ values were identical in the first 175 and the last 125 specimens examined.

To further illustrate the pattern and extent of discordant readings, we plotted a histogram of the number of categories of disagreement and the direction for Observer A relative to Observer B (Fig. 1) (9). The data show that 79 of 96 (82%) disagreements involve only one category with a tendency for Observer A to assign higher scores of one-category change only with respect to Observer B. Larger disagreements were evenly divided between relative under- and overscoring. Most of the one-category overscoring was accounted for by the assignment of "mild atypia" by Observer A in specimens labeled as "normal metaplasia" by Observer B.

To determine whether there was a consistent pattern of disagreement between observers, we had a third observer, a trained cytopathologist, review independently all samples in which disagreement occurred. Of the 96 discordant specimens, the reference pathologist agreed with Observer A in 44 cases, with Observer B in 39 cases, and with neither in 13 cases.

Within Observer Agreement. For this analysis we selected 60 representative specimens from 49 subjects. A five-category scale was used that excluded "no metaplasia" and "malignant change." By this time the cytotechnologists had agreed that the absence of metaplasia was extremely rare and that they were unsure that this occurred within satisfactory specimens. It was agreed by consensus to classify these specimens subsequently in the unsatisfactory category. Sixty specimens were chosen so that $\kappa$ would be normally distributed, which occurs when $n > 2$ (number of categories)$^2$ (10). Table 2 displays the within observer agreement matrices for observers A and B, who recorded diagnoses on two separate occasions at least 2 wk apart, to minimize recall. For Observer A, there was 80% (48 of 60) complete agreement, ($\kappa = 0.73$) while Observer B performed less well (agreement, 62%; $\kappa = 0.49$). The difference in $\kappa$ values in significant at $P < 0.04$.

The histograms in Fig. 2, a and b, illustrate the extent and pattern of disagreement for the second reading relative to the first for the two observers. For Observer A, 7 of the 12 discordant readings were within one category, and overall this observer tended to overscore on the second reading relative to the first. Observer B also tended to "overscore" the second time, but of the 23 discordant readings, 18 were within one category only. Overall, for the two observers, 71% (25 of 35) of the within observer disagreements were within one category only.

Initially, we specified that a "response" to treatment would be defined as at least a two-category change in atypia for subjects presenting at the start of the study with either "moderate" or "marked atypia," and a change to at least "normal metaplasia without atypia" for subjects presenting with "mild atypia." If we were to retain these rules (and eliminate unsatisfactory readings since these would be repeated), then for Observer A there would be 5 erroneous "responses" among 46 judgments (11%), and for Observer B, there would be 11 erroneous "responses" among 51 judgments (22%) (Table 2). Most of these 16 errors are accounted for by disagreement between the categories "mild atypia" and "normal metaplasia" (11 of 16) which represents a single category change.

DISCUSSION

General Discussion This study of between and within observer agreement was prompted by the need to evaluate the reproducibility of the diagnostic criteria for the primary outcome measure, sputum cytology, in a chemoprevention trial. The level of agreement obtained in this study is impressive compared with others in the medical literature where subjective judgments have been evaluated (11). Moreover, the performance of the cytotechnologists in this study is superior to that reported in a similar chemoprevention trial of folate and vitamin B$_2$ in smokers, where complete intraobserver agreement was obtained in only 22 of 40 (55%) specimens (12). Despite the variability observed in the reported trial, it had sufficient power to detect a beneficial effect in the treatment group compared with the placebo control.

Other studies have addressed problems of disagreement in sputum cytology from patients diagnosed with lung cancer (13-18). Most of these were concerned with determining sources of variability in diagnosis, or with comparisons of cytology results obtained from subjects with known clinical outcomes. An excellent study of between observer agreement across several institutions has been reported; however, the arbitrary designation of the host institution diagnosis as a "gold standard" preempted evaluation of within observer reproducibility (16). None of these studies formally evaluated agreement within observers nor was agreement beyond chance accounted for, except for an evaluation of observer variability in the histopath-

### Table 1 Interobserver agreement matrix

<table>
<thead>
<tr>
<th>Observer A</th>
<th>Unsatisfactory</th>
<th>No metaplasia</th>
<th>Normal metaplasia</th>
<th>Metaplasia, mild atypia</th>
<th>Metaplasia, moderate atypia</th>
<th>Metaplasia, marked atypia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsatisfactory</td>
<td>34</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No metaplasia</td>
<td>3</td>
<td>3</td>
<td>52</td>
<td>26</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Normal metaplasia</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>54</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>Mild atypia</td>
<td>3</td>
<td>1</td>
<td>19</td>
<td>60</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Moderate atypia</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marked atypia</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Number = 300; agreement = 68% (204 of 300); $\kappa = 0.58$. 

1217
SCREENING PHASE DATA AND EVALUATION OF OBSERVER VARIATION

Fig. 1. Histogram of direction and extent of disagreements for Observer A relative to Observer B.

Table 2 Intraobserver agreement matrix

<table>
<thead>
<tr>
<th>Second reading</th>
<th>Normal</th>
<th>Mild atypia</th>
<th>Moderate atypia</th>
<th>Marked atypia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer A, first reading*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsatisfactory</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Normal metaplasia</td>
<td>14</td>
<td>1</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>Mild atypia</td>
<td>1</td>
<td>19</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Moderate atypia</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Marked atypia</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Observer B, first reading*

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Mild atypia</th>
<th>Moderate atypia</th>
<th>Marked atypia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsatisfactory</td>
<td>7</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal metaplasia</td>
<td>1</td>
<td>14</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Mild atypia</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Moderate atypia</td>
<td>10</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marked atypia</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Agreement = 80% (48 of 60); \( \kappa = 0.73 \).
* Agreement = 62% (37 of 60); \( \kappa = 0.49 \).

The data reported here evaluate the reproducibility but not the validity of the scoring decisions made. A gold standard does not exist currently for the diagnostic classification used in this study. One could arbitrarily designate an expert opinion as a gold standard based on experience and credibility, but we are reluctant to do so. Our primary concern is to minimize both bias and imprecision in the assessment of outcomes and to determine whether treatment with etretinate produces reversal in abnormal sputum cytology despite the variability associated with the measurement.

The study reported here indicates that sputum cytology is sufficiently credible and reproducible to be used as an outcome basis to adjudicate specimens for which there is disagreement between the mild atypia and normal metaplasia categories, since this accounts for the majority of erroneous response judgments. It would also be reasonable to consider restricting the analyses for response to those with moderate and severe atypia only, thus eliminating most of the miscategorizations. However, this could be done only if the study retained sufficient power to detect a reasonable effect size.

We are encouraged by the results of this study which have taught us a great deal about the outcome measure for our trial. Despite the variability observed in sputum diagnosis, the performance of these observers is superior to that reported in a similar study involving fewer subjects where a significant difference was detected using the same outcome measure (12). Thus, we are confident that, using the rules specified originally together with refinement of reporting using the most consistent observer, we will have the power to detect important differences between treatments if they exist. We will explore the screening phase data further before making a final decision on the precise methods for comparing the treatment groups.
measure in our chemoprevention trial. Results from the screening phase of such a trial may be useful in refining outcome measures and can avoid manipulation of the observed trial results.

An important final consideration of this approach will be how the diagnostic categories are distributed in the final trial results relative to the screening phase, since $\kappa$ is a prevalence-dependent statistic (20). Therefore, conclusions about reproducibility based on screening phase data may be premature until the final data are available.

REFERENCES


Use of Screening Phase Data to Evaluate Observer Variation of Sputum Cytodiagnosis as an Outcome Measure in a Chemoprevention Trial

George P. Browman, Andrew Arnold, Mark N. Levine, et al.


Updated version  Access the most recent version of this article at:  
http://cancerres.aacrjournals.org/content/50/4/1216

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.