AutoPap System Performance in Screening for Low Prevalence and Small Cell Abnormalities

James S. J. Lee, Ph.D., Paul Wilhelm, M.S., Leonard Kuan, M.S., Dayle G. Ellison, M.S., Xingye Lei, Ph.D., Seho Oh, Ph.D., and Stanley F. Patten, Jr., M.D., Ph.D., F.I.A.C.

OBJECTIVE: To summarize the design principles of the AutoPap System evaluation score by evaluating slides having a low prevalence of abnormal cells and small cell abnormalities and assessing the evaluation score as a diagnostic tool.

STUDY DESIGN: Data from two clinical studies conducted using the AutoPap System and data obtained from the evaluation score training slides were analyzed to demonstrate the effectiveness of the evaluation score. The clinical studies included a prospective, intended-use study involving approximately 13,000 slides and a comprehensive sensitivity study using approximately 1,200 slides from five laboratories. The evaluation score training set consisted of 4,174 slides from 10 laboratories.

RESULTS: The robust design of the AutoPap evaluation score was demonstrated by similar detection capabilities and sensitivities to slides having either a low or high prevalence of abnormal cells. No significant difference in performance was detected between the small cell slides and the comparison groups of carcinoma in situ and invasive squamous carcinoma having normal-sized abnormal cells. In addition, the evaluation scores corresponded well to the diagnostic severity of the slides. (Acta Cytol 1997;41:56–64)

Keywords: cervical smears, mass screening, AutoPap.

Abnormal biologic change may be found in a number of cellular formations, such as free-lying cells and cellular aggregates (sheets, syncytia and clusters).1 The evaluation score relies upon the detection of every form of abnormality. To enhance detection of an apparently rare event (for example, a single truly abnormal cell), the evaluation score integrates the results of three image interpretation algorithms to detect cellular and cell population evidence of abnormality: the single cell algorithm (for free-lying cells), a group algorithm (for aggregate sheets and syncytia) and a thick group algorithm (for aggregate syncytia and clusters).

For each slide successfully processed, the AutoPap System2-8 generates an evaluation score ranging in value from 0 to 1.0. This evaluation score re-
fects the likelihood that a slide is abnormal: the higher the evaluation score, the more likely that abnormal cells are present. By comparing this score with a preset threshold, the AutoPap System (hereafter called the device) determines whether the slide needs to be reviewed by a cytotechnologist.

To confirm the effectiveness in a primary screening application and in accordance with the study design, we assessed the evaluation score with slides having a low prevalence of abnormal cells and small cell abnormalities. In addition, correlation of the evaluation score value with the diagnostic severity of the slides was assessed.

**Evaluation Score Design Principles**

The challenge in designing the evaluation score was to ensure that detected cellular changes are due to the measurement of abnormal cellular changes and not to other factors. That is, an appropriate balance between sensitivity to abnormal cells and specificity to artifacts and normal cells must be obtained. This process can be referred to as maximizing the classification signal-to-noise ratio. (*Signal* is the detection of abnormality, and *noise* is normal material incorrectly classified as abnormal objects.)

The evaluation score was designed to maximize the classification signal-to-noise ratio using a confidence-based information integration method. Noise is caused, in part, by uncertainties in cell appearance, in locating cell nuclei, in obtaining optimum focus on a cell and in measuring nuclear size and stain. If the certainty at each stage of the system is recorded, an overall estimate of classification confidence can be made. Such confidence estimates help reduce noise by quantifying the information content and reliability of each classification result. Thus, each object (cell, aggregate or artifact) has not only a classification result that reflects an outcome but also an associated confidence value. Thus, the evaluation score integrates information with heavy reliance upon confidence information for noise reduction.

Noise is also reduced by normalizing variations caused by known sources, such as variations in specimen collection and specimen preparation. When calculating the evaluation score, these variations are normalized in the following ways:

- **Reference Cell Feature Measures for Feature Calibration**
  Intermediate squamous cell nuclei are a standard (reference) against which other cells can be compared. The image interpretation algorithms use information about size, texture and stain quality of detected reference cell nuclei as a means of normalizing stain and fixation variation from specimen to specimen.

- **Evaluation Score Architecture**
  The evaluation score algorithm architecture employs a “divide-and-conquer” strategy that includes multiple classifiers, each of which handles a specific range of specimen preparation. This allows normalization of major specimen variation or bias due to collection, fixation and preparation variation.

  To enhance signal quality, cellular contextual information is considered. That is, abnormal objects tend to be closely associated (proximal) rather than randomly distributed over the slide. The object classification results of the image interpretation algorithms are compared, and when several objects classified as potentially abnormal are found in close proximity, the objects are treated as more likely to be truly abnormal.

**Algorithm Architecture**

Calculation of the evaluation score is a three-step process consisting of 4× processing, 20× processing and calculation of the final evaluation score.

- **4× Processing**
  The AutoPap System first assesses all areas on an individual slide using the 4× processing algorithm (Figure 1). A 4× image is then acquired using the 4× objective lens. The pixel size is 2.75×2.75 μm.

  This initial algorithm assessment produces the road map, or guidelines, used by subsequent algorithms in finding and measuring the information on the slide. The 4× algorithm processes the information on a slide in three major steps:

  1. **Image Segmentation**
     Identifies areas of small objects, large objects and
p. 58

Acta Cytologica

Lee et al

potential cell groups and passes them to a feature computation module.

2. Feature Computation

Measures properties of objects and image regions that were detected during image segmentation.

3. Ranking

Derives two ranks from the computed features: one for the abnormal free-lying cells (squamous intraepithelial lesion [SIL] rank) and one for cellular aggregates (group rank). The ranks reflect the likelihood of a 20x field of view containing an SIL or a group. These rank values, as well as bubble edge flags and three detected object histograms, are passed to the 20x algorithms for further processing.

20x Processing

After completing 4x algorithm processing, the information contained on the slide is subjected to a second level of review and quantitative assessment by the 20x algorithms (Figure 2). The 20x image is acquired based on a 0.75 NA 20x objective lens. The pixel size is 0.55x0.55 μm.

During this second level of processing, fields of view selected as a result of 4x processing are captured at high magnification (20x) and analyzed. Each image is assessed by the five algorithms identified in Figure 2. (Note that two of these algorithms provide imaging quality information rather than abnormality analysis information: the stripe detection algorithm and the focus check algorithm.) The compiled information from each image is next combined into a single results structure that encompasses the 20x results. The algorithms then assess the likelihood of abnormality within the context of the following features that contribute to each image:

- Classification of free-lying cells
- Stain characteristics of the cells
- Confidence levels of the classification results
- Characteristics of each cellular aggregate
- Other contextual information from multiple features
- Image quality for each processed image

Calculating the Slide Evaluation Score

The evaluation score architecture consists of four parallel scoring classifiers and one preparation membership classifier. Each of these four scoring classifiers is trained to be most effective for a specific range of slide preparations. A fifth classifier, the membership classifier, calculates a similarity measure of the preparation of the slide being scored to the preparation type for which each scoring classifier was developed (Figure 3).

Each of the classifiers uses a subset of slide features to calculate classifier results. The slide features include cellular features derived from isolated cells, cellular aggregates and proximal cells. In addition, specimen preparation-related features (stain and fixation qualities) are used to normalize the cellular features for robust classification and to achieve the highest possible classification signal-to-noise ratio.
When calculating the evaluation score, the scores are combined from the four scoring classifiers weighted by the similarity measure calculated by the membership classifier. This architecture allows normalization in scoring of specimen collection and preparation variation or bias, which may influence the image interpretation algorithms. This architecture is designed to yield preparation-independent scoring results.

**Performance Evaluation**

To confirm the effectiveness of the evaluation score, its performance is evaluated for slides having a low prevalence (small number) of abnormal cells. In addition, the effectiveness of the evaluation score in detecting small cell carcinoma *in situ* (CIS) (CIN III) and small cell invasive squamous carcinoma cervical/vaginal smear slides was determined. Finally, the effectiveness of the evaluation score as a diagnostic tool when using the AutoPap System as a primary screener was assessed.

Data from clinical studies of the AutoPap System\(^9\text{-}^{11}\) and evaluation score training set were used for the evaluation.

**Sensitivity to Low Prevalence Slides**

Analysis was conducted on 4,174 slides from the slide evaluation score training set to evaluate the effectiveness of the AutoPap evaluation score in classifying slides having a low prevalence of abnormal
cells. Of the 4,174 slides selected for use, 2,293 slides were abnormal.

Of the 2,293 abnormal slides, the score properly classified 2,112 slides as Review at a 50% sort rate. These slides were removed from this analysis as “less difficult” or “easily detectable.” The remaining 181 abnormal slides are considered “difficult” abnormal slides and were used as the study set for this analysis.

Each of these 181 slides was then rescreened by the NeoPath cytopathology department to determine the actual number of abnormal cells on each slide. Each slide was classified with both a 10-cell and 20-cell count limit. Slides with abnormal cell counts in excess of 20 were classified as normal prevalence. Based on these manual cell counts, the slides were categorized as follows:

- 20-Cell cutoff limit
- Difficult normal prevalence (109 slides with ≥ 20 abnormal cells)
- Difficult low prevalence (72 slides with < 20 abnormal cells)
- 10-Cell cutoff limit
- Difficult normal prevalence (158 slides with ≥ 10 abnormal cells)
- Difficult low prevalence (23 slides with < 10 abnormal cells)

Table I shows the sensitivity of the device in detecting these slides at a 30% sort rate. Sensitivity is based upon the ability of the score to classify abnormal slides as Review. Data are presented for slides sorted at a low prevalence rate of < 20 abnormal cells and at a low prevalence rate of < 10 abnormal cells per slide.

Note that Table I does not depict the true distribution results when using the AutoPap System as an automated screener for all abnormal slides in this population. True distribution results for the entire abnormal slide population are shown in Table II. Note that the overall evaluation score sensitivity at a 30% sort rate is 96.9%.

Fisher’s exact test was used to test the association between prevalence and the AutoPap System evaluation score sensitivity. For both sets of abnormal slides containing < 20 diagnostic cells or < 10 diagnostic cells, the P values were close to 1.0. This result indicates that the evaluation score demonstrated an equivalent sensitivity level in the detection of abnormality on low prevalence slides (at < 20 cell count and < 10 cell count) as compared to normal prevalence slides (≥ 20 cells).

Additionally, the analysis showed no significant difference in the sensitivity of the device between the two groups of difficult abnormal slides. This demonstrates that the evaluation score slide classification sensitivity to abnormality (at both a 20 and a 10 abnormal cell count limit) was independent of (and not dependent on) the actual count of abnormal cells.

<table>
<thead>
<tr>
<th>Group</th>
<th>Review</th>
<th>No review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficult low prevalence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(&lt;20, n = 72)</td>
<td>61.1% (44/72)</td>
<td>38.9% (28/72)</td>
</tr>
<tr>
<td>Difficult normal prevalence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(≥20, n = 109)</td>
<td>61.5% (67/109)</td>
<td>38.5% (42/109)</td>
</tr>
<tr>
<td>Difficult low prevalence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(&lt;10, n = 23)</td>
<td>60.9% (14/23)</td>
<td>39.1% (9/23)</td>
</tr>
<tr>
<td>Difficult normal prevalence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(≥10, n = 158)</td>
<td>61.4% (97/158)</td>
<td>38.6% (61/158)</td>
</tr>
</tbody>
</table>

Sort rate = 30%, n = 181.

Table I Classification Results of Low Prevalence vs. Normal Prevalence Slides
AutoPap System Performance
Sensitivity to Slides Showing Small Cell Abnormalities

An analysis was conducted on abnormal slides to determine the sensitivity of the AutoPap System evaluation score to small cell CIS and small cell invasive squamous carcinoma at a 30% sort rate.

Table III compares the device sensitivity rates for the small cell slides from these two diagnostic categories with the sensitivity rates for these two categories when small cell slides were excluded. (Table III provides the device detection rate on the first line, the 95% confidence intervals on the second line and the number of slides detected/scanned on the third line.)

Fisher's exact test was used to test the sensitivity difference between the small cell groups and non–small cell abnormal slide groups. The P value was >0.99 for the CIS (CIN III) group and 0.145 for the invasive squamous carcinoma slide group. Since neither P value was less than .05 (5% significance level), there was no demonstrated difference between sensitivities or performance of the device in detecting small cell and non–small cell slides for the diagnostic categories studied.

Note that the 95% confidence intervals for the sensitivity of the device to small cell slides and for the non–small cell slides overlap, indicating that any difference in the performance of the device between these groups of slides is not statistically significant.

Evaluation Score Correlation with Disease Severity

The evaluation score reflects the likelihood that a slide contains abnormal cells in any disease category. The higher the score, the more likely the slide is to contain abnormal cells. It is desirable that the device identify those slides with a severe lesion as having a higher likelihood of abnormality than those slides with a less severe lesion.

The following two data sets were analyzed to evaluate the correlation between evaluation score and the severity of disease:

- The sensitivity study data set derived from the premarket approval (PMA) clinical data for the quality control application of the device. This study, referred to as the historical sensitivity study (HSS), consisted of slides pulled from laboratory archives over a two-year period. Selecting slides in this manner resulted in a large sample size for each diagnostic category studied (Figure 4).

- The clinical data derived from a prospective, intended-use study of the device as a primary screener. These data were reported separately in the PMA filing for the device as the clinical evaluation study (CES) and the current archive study (CAS).

The CES report analyzed the device results for slides classified as within normal limits during the initial screening and also the ability of the device to detect false negative slides. The CAS report assessed the sensitivity of the device to the abnormal

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Small cell CIS slides (CIN III)</th>
<th>Other types of HSIL</th>
<th>Small cell invasive squamous carcinoma</th>
<th>Other types of squamous carcinoma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity at 30%</td>
<td>100.0%</td>
<td>99.0%</td>
<td>92.9%</td>
<td>99.4%</td>
</tr>
<tr>
<td>sort rate</td>
<td>81.5–100</td>
<td>97.9–99.5</td>
<td>68.3–99.6</td>
<td>96.9–99.9</td>
</tr>
<tr>
<td></td>
<td>(18/18)</td>
<td>(616/622)</td>
<td>(13/14)</td>
<td>(172/173)</td>
</tr>
</tbody>
</table>

Table III: Sensitivity of AutoPap System to Small Cell CIS and Small Cell Invasive Squamous Carcinoma Slides as Compared with Other CIS and Squamous Carcinoma Slides
Figure 4 shows the evaluation score box-and-whisker plots of the two data sets (HSS and CES/CAS) for each diagnostic category studied. The middle line corresponds to the median value, and the solid boxes correspond to the upper quartile (75%) and lower quartile (25%) of evaluation scores. In addition, the upper extreme (excluding outliers) and lower extreme (excluding outliers) for

slides that occurred within a normal workload (Figure 5).
each box are plotted. The extreme is either true maximum (minimum) or the point that is farther than 1.5 times the interquartile range from the upper (lower) quartile.

As shown in Figure 4, cancer and high grade SIL (HSIL) slides are distributed on the higher end of
the evaluation score, and normal slides are distributed over the lower end of the score. As expected, atypical squamous cells of undetermined significance/atypical glands of undetermined significance (ASCUS/AGUS) and low grade SIL (LSIL) slides are distributed in the middle range of the evaluation scores.

Figure 5 shows the plots of the evaluation score distribution for the slides from the two data sets (HSS and CES/CAS) for each diagnostic category studied. Note that the median evaluation score values increase monotonically for normal, ASCUS/AGUS, LSIL, HSIL and cancer slides, indicating that the evaluation scores produced by the device correspond well to the diagnostic severity of the slides.

Discussion

The AutoPap evaluation score is designed to maximize the classification signal-to-noise ratio. It integrates results of three image interpretation algorithms to detect cellular and cell population evidence of abnormality using a confidence-based information integration method. This design principle provides the foundation for highly accurate classification performance.

The evaluation score is developed using a comprehensive training strategy that incorporates a large number of abnormal slides from different diagnostic categories as well as slides representing the subcategories within each group. In addition, the algorithm design includes a confidence-based object classification strategy, which produces evaluation scores that appear to correlate well with the severity of disease.

The AutoPap System supports a high-resolution and highly stable optical imaging system of 0.55 × 0.55-μm pixel size. Based on the morphometric characteristics of small cell abnormalities— that is, 52–80 μm² (172–264 pixels) of nuclear area, the imaging resolution can precisely measure small nuclei and detect these abnormal cells.

Low prevalence and small cells present challenges to semiautomated cytology systems. The design of the AutoPap System evaluation score maximizes the classification signal-to-noise ratio, and therefore the device is able to detect slides having a low prevalence of abnormal cells. The AutoPap System evaluation score demonstrates similar detection capabilities and sensitivities to slides having either a low or high prevalence of abnormal cells.

The results presented here validate the effectiveness of the evaluation score in detecting abnormal slides containing specific small cell abnormalities. The results of this analysis suggest that the AutoPap System evaluation score shows promise as a diagnostic tool in the assessment of cervical/vaginal smear slides.

References