Evaluation of quantitative image analysis criteria for the high-resolution microendoscopic detection of neoplasia in Barrett’s esophagus

Timothy J. Muldoon  
Nadhi Thekkek  
Darren Roblyer  

Rice University  
Department of Bioengineering  
6100 Main Street  
Houston, Texas 77005

Dipen Maru  
The University of Texas  
M.D. Anderson Cancer Center  
Department of Pathology  
1515 Holcombe Boulevard  
Houston, Texas 77030

Noam Harpaz  
Mount Sinai Medical Center  
Division of Pathology  
1 Gustave L. Levy Place  
New York, New York 10029

Jonathan Potack  
Sharmila Anandasabapathy  
Mount Sinai Medical Center  
Division of Gastroenterology  
1 Gustave L. Levy Place  
New York, New York 10029

Rebecca Richards-Kortum  
Rice University  
Department of Bioengineering  
6100 Main Street  
Houston, Texas 77005

1 Introduction

The incidence of esophageal adenocarcinoma (EAC) has increased sixfold in the United States over the past 30 years. The current method of surveillance of Barrett’s esophagus has one of the poorest overall five-year survival rates of all cancers, with reported five-year survival rates of 14% and 0% for stage III and stage IV disease, respectively. Surgical treatment for locally advanced EAC carries significant morbidity and mortality. Early detection of disease is vital to improve long-term survival rates, facilitate endoscopic therapy, and improve the overall quality of life of affected patients.

The current method of surveillance of Barrett’s esophagus involves endoscopic white-light examination with four-quadrant biopsy, a procedure that has been shown to miss neoplasia in up to 57% of cases. Because of these limitations, high-resolution technologies that can serve as an adjunct to white-light endoscopy have been proposed to increase diagnostic accuracy for the detection of high-grade dysplasia (HGD) or early cancer. These technologies include confocal microscopy and endocytoscopy. Of these, confocal endomicroscopy has shown the highest accuracy to date based on single-center evaluation with an experienced endomicroscopist interpreting the in vivo images. Unfortunately, such real-time interpretation requires considerable training and is subject to intraobserver variability. For broader clinical application, particularly outside a university setting, the application of image-analysis software using well-defined image classification algorithms has the potential to reduce subjectivity and improve diagnostic accuracy.
Quantitative image analysis, based on computational analysis of textural features within a digital image, offers an objective means to improve the consistency and speed of diagnosis at the point of care. Image texture features, based on the spatial distribution and organization of pixels, can be computed rapidly. Textural features, such as entropy, frequency content, and pixel pair correlation values, may not be apparent to a human observer, but they relate important information about the structure within an image. Images from a dysplastic region, for example, would be more likely to have a lower pixel pair correlation value and a higher entropy value because of the decreasing orderly arrangement of cells (and therefore corresponding image pixels) when compared to normal glandular tissues. Since there are a large number of easily calculated textural features, searching many features at once can identify a single feature or pair of features that enable high-fidelity classification of the image data. By limiting the number of features calculated at the time of image acquisition, rapid diagnostics are possible at the point of care. With the increasing availability and power of inexpensive and portable computers, on-site quantitative analysis and classification techniques could be applied to a wide array of high-resolution imaging schemes.

To extend the benefits of high-resolution imaging to a broader patient population, and to create a more objective means of evaluating optical biopsy data, we developed a quantitative image analysis and classification algorithm to analyze data from a high-resolution microendoscope (HRME) device. The HRME is capable of producing images of the esophageal mucosa at subcellular resolution without the need for expensive optics or scanning electronics. Indeed, the total cost of the HRME components is approximately $2500. In a previous proof-of-principle study, we demonstrated that images acquired with the HRME can distinguish neoplastic [high-grade dysplasia (HGD) or cancer] from non-neoplastic [intestinal metaplasia or low-grade dysplasia (LGD)], Barrett’s mucosa based on qualitative image assessment.11

The goal of this study was to develop quantitative image analysis criteria for HRME images to discriminate neoplastic from non-neoplastic esophageal mucosa. We compared the accuracy of discrimination achieved using visual interpretation by experienced pathologists or gastroenterologists trained in the analysis of the optical images to that achieved with computer-aided classification algorithms based on quantitative analysis of image textural features. Histopathology was considered the gold standard.

2 Methods

2.1 Patients

Patients over 18 years of age with a previous diagnosis of Barrett’s esophagus were asked to participate in the study. Informed consent was obtained from all study participants, and the study was reviewed and approved by the Institutional Review Boards at the University of Texas M. D. Anderson Cancer Center, Rice University, and The Mount Sinai Medical Center. Subjects underwent conventional endoscopy with standard four-quadrant biopsy surveillance; a subset of patients (those with endoscopically suspected neoplasia) underwent endoscopic mucosal resection (EMR). Following resection, biopsies or EMR specimens were imaged with the HRME device.

A solution of proflavine (Sigma-Aldrich) dissolved in water at a concentration of 0.01% (w/v) was prepared prior to performing imaging. The contrast agent solution was directly applied to the epithelial surface of the resected tissue with a dropper, and imaging with the HRME device was performed immediately. The application of proflavine does not discolor the tissue surface, and is not detectable in tissue slides prepared using standard histopathology processing and hematoxylin and eosin (H&E) staining (Fig. 1).

After imaging, the tissue was returned for standard histopathology processing, and slides were later reviewed by a single, expert gastrointestinal pathologist (D.M.) blinded to the results of the HRME imaging. Each measurement site used in this study was correlated to a histopathology-confirmed diagnostic category: Barrett’s intestinal metaplasia (IM), low-grade dysplasia (LGD), high-grade dysplasia (HGD), or esophageal adenocarcinoma (EAC).

2.2 Imaging System

The high-resolution microendoscope (HRME) device has been previously described in detail.12 Briefly, images are acquired with this device by placing the tip of the fiber bundle image guide into direct contact with the epithelial surface of the tissue. Excitation light from a blue LED with a center wavelength of 455 nm is delivered through the fiber bundle. The fluorescence emission from the topically applied fluorescent contrast agent, proflavine, is collected through the fiber bundle and focused onto a CCD camera, and a digital image is stored for future processing and analysis. The HRME system has a circular field of view with a diameter of 750 microns; the lateral spatial resolution of the system is approximately 4 microns, and images are displayed at 4 frames per second.

2.3 HRME Image Analysis

Digital HRME images were reviewed to determine whether the endoscope tip was in contact with the tissue surface or whether the probe tip moved during image acquisition. Images showing such artifacts were discarded and not used in subsequent analyses.

Digital HRME images were reviewed by two expert pathologists and two expert gastroenterologists, already familiar with microendoscopy. Prior to reviewing the entire set of images, reviewers were shown a subset of 16 images, labeled with the corresponding histopathologic diagnosis of IM, LGD, HGD, and EAC. This training set included 8 images collected from sites with a diagnosis of IM or LGD (non-neoplastic), and 8 images from sites diagnosed with neoplasia (HGD or EAC). Reviewers were then shown the complete set of images in a randomized order and asked to score each image as either “neoplastic” or “non-neoplastic,” where “non-neoplastic” corresponds to a pathologic diagnosis of IM or LGD, and “neoplastic” corresponds to HGD or EAC. Results of visual image interpretation were compared to the histopathology-confirmed diagnosis at each site; sensitivity and specificity were calculated for each observer.
2.4 Image Classification Algorithm

In addition to subjective image interpretation, we explored the diagnostic ability of quantitative image analysis. For each image, 59 distinct features were computed (Table 1). First-order statistical features (variance, entropy, etc.) were calculated directly from the raw pixel values. A gray-level co-occurrence matrix (GLCM) with pixel offsets ranging from 1 to 10 was used to calculate additional textural feature groups (correlation, contrast, homogeneity, and energy). Each GLCM feature group contained 10 distinct features, corresponding to each pixel offset. To detect nuclear features, an extended rectangular tessellation was calculated from the centroids of the nuclear features to calculate inter-nuclear distances. A Fourier transform was applied to the image. Voronoi tessellations were calculated from the centroids of the nuclear features to calculate inter-nuclear distances. A Fourier transform was applied to each image to calculate the power spectrum; this was divided into 10 partitions to represent the frequency components of the image. The contribution of each partition is represented as a fraction of the total power spectrum.

A diagnostic algorithm was developed to classify each image as non-neoplastic or neoplastic using these image features as input. Histopathology again was used as the gold standard; sites with a pathologic diagnosis of Barrett’s metaplasia or Barrett’s metaplasia with low-grade dysplasia were considered to be non-neoplastic, while sites with a pathologic diagnosis of Barrett’s metaplasia with high-grade dysplasia or esophageal adenocarcinoma were considered to be neoplastic. The classifier was based on two-class linear discriminant analysis; a sequential forward feature-selection algorithm initially was used to select the best performing subset of up to 10 image features to classify the data. Initially, the best performing single feature was identified, and then subsequent features were selected that gave maximum performance when combined with previously selected features. The algorithm was developed using fivefold cross-validation; each measurement site was initially randomly assigned to one of five groups. Four-fifths of the data were then used to train the linear classification algorithm, and the remaining one-fifth of the data were used to test the algorithm. This cycle was repeated four additional times so that the algorithm was tested using data from each site. Performance was monitored by calculating the area under the curve (AUC) of the classifier.

Alternatively, a three-class diagnostic algorithm was developed to classify each image as: (1) Barrett’s intestinal metaplasia or Barrett’s low-grade dysplasia, (2) Barrett’s high-grade dysplasia, and (3) adenocarcinoma. A categorical tree-based classifier was used. This algorithm was allowed to choose inputs from the entire feature set. The tree-based classifier was pruned to three terminal nodes to avoid overfitting; two features were selected to perform this step. The predicted classification results from each classifier were then compared to the actual histopathology for each site.

3 Results

3.1 Patients and Measurement Sites

Nine subjects were enrolled in the study. Endoscopic mucosal resection (EMR) specimens were obtained from six of these patients, and biopsy specimens were obtained from the remaining three patients. Images were obtained from 139 unique sites; images from 128 of these sites passed the quality control (QC) review and were used for further analysis (Table 2). Figure 1 shows representative HRME images (top row) and corresponding histopathology (bottom row) of metaplasia/LGD, HGD, and EAC, respectively. All images are at the same scale for comparison. Large, well-organized glands can be seen in the metaplasia/LGD Barrett’s case, while smaller glands with disrupted borders are visible in the
HGD case. The image of EAC shows extreme disruption of glandular organization and crowded, abnormal cells.

### 3.2 Subjective Image Interpretation

Subjective scoring of the images by two expert gastroenterologists achieved an average sensitivity of 87% and an average specificity of 53% for neoplasia, with a kappa statistic of 0.25, indicating fair agreement. The two expert pathologists achieved an average sensitivity of 87% and an average specificity of 68% for neoplasia, with a kappa statistic of 0.27, indicating fair agreement. The combined average performance of all four reviewers was 87% sensitivity and 61% specificity.

### 3.3 Quantitative Classification Algorithm

Quantitative image features were ranked according to their diagnostic ability. Image features in Table 1 are listed in order of descending area under the receiver operating characteristic (ROC) curve for the single-feature, two-class linear discriminant analysis classifier. The single best performing feature was found to be the GLCM correlation with a pixel offset of 10. Figure 2(a) depicts a box plot illustrating the average GLCM correlation values for samples diagnosed histologically as non-neoplastic samples (IM or LGD), HGD, and HGD case. The image of EAC shows extreme disruption of glandular organization and crowded, abnormal cells.

#### Table 1

Quantitative image features. Image features are listed in decreasing order of diagnostic performance when used as a single input feature for linear discriminant analysis.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation (10 features)</td>
<td>Pixel neighbor correlation over the entire image</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>Standard deviation of grayscale values</td>
</tr>
<tr>
<td>Variance</td>
<td>Variance of pixel grayscale values</td>
</tr>
<tr>
<td>Energy (10 features)</td>
<td>Sum of squares in gray-level co-occurrence matrix (GLCM)</td>
</tr>
<tr>
<td>Frequency (10 features)</td>
<td>Frequency distribution of pixel values</td>
</tr>
<tr>
<td>Entropy</td>
<td>Statistical measurement of randomness of grayscale values</td>
</tr>
<tr>
<td>Mean nuclear separation</td>
<td>Mean nuclear separation as calculated by Voronoi tessellation</td>
</tr>
<tr>
<td>Std. dev. nuclear separation</td>
<td>Standard deviation of nuclear separation as calculated by Voronoi tessellation</td>
</tr>
<tr>
<td>Nuclei per unit area</td>
<td>Number of nuclei detected divided by area of region of interest</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>Measure of the flatness of the pixel value distributions</td>
</tr>
<tr>
<td>Skewness</td>
<td>Measure of the symmetry of the pixel value distribution</td>
</tr>
<tr>
<td>Contrast (10 features)</td>
<td>Measure of pixel intensity compared to its neighbors over the entire image</td>
</tr>
<tr>
<td>Homogeneity (10 features)</td>
<td>Closeness of the distribution of the GLCM elements to the diagonal</td>
</tr>
<tr>
<td>Mean minimum nuclear separation distance</td>
<td>Average minimum nuclear separation as calculated by Voronoi tessellation</td>
</tr>
</tbody>
</table>

#### Table 2

Number of sites imaged by histologic diagnosis.

<table>
<thead>
<tr>
<th>Diagnostic category</th>
<th>Histologic diagnosis</th>
<th>Number of sites imaged</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-neoplastic</td>
<td>IM</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>LGD</td>
<td>53</td>
</tr>
<tr>
<td>Neoplastic</td>
<td>HGD</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>EAC</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>128</td>
</tr>
</tbody>
</table>

**Fig. 2** (a) Box and whiskers plot showing the GLCM correlation by diagnostic category. The central line in each box represents the mean, while the top and bottom edges represent the 25th and 75th percentile. The notches in the boxes represent a 95% confidence interval for the mean. (b) Features values for top two performing features by measurement site. Non-neoplastic sites are plotted as squares; neoplastic sites are plotted as crosses. The decision line is shown.
EAC. On average, the GLCM correlation level is lower in sites with HGD or EAC when compared to non-neoplastic sites.

The top performing combination of two features was found to be the GLCM correlation value at an offset of 10, and the frequency contribution at an offset of 6; this is the relative contribution to the total power spectrum coming from the frequencies over the sixth partition from the Fourier transform. Figure 2(b) shows a scatter plot of these two features for each of the 128 sites in the data set. Non-neoplastic samples, with a pathologic diagnosis of Barrett’s with or without LGD, are shown as squares; neoplastic samples, with a pathologic diagnosis of HGD or EAC, are shown as crosses. The decision line associated with the two-class linear discriminant classifier is shown as a straight line on the plot.

Figure 3(a) shows the area under the curve as a function of the number of image features selected in the linear discriminant algorithm. The AUC increases from one to two features and then reaches a plateau. Figure 3(b) shows a scatter plot of the posterior probability that each site is neoplastic (HGD or EAC) as calculated by the linear classifier using two input features; samples are grouped by histopathologic diagnosis. Figure 3(c) shows the ROC curve for the linear discriminant classifier based on these two features. The AUC is 0.92, and the sensitivity and specificity at the Q-point are 87% and 85%, respectively.

A tree-based classifier was developed to classify samples into three groups (non-neoplastic, HGD, and cancer) using automated feature selection. The features chosen by this algorithm were again GLCM correlation and frequency contribution. Table 3 summarizes the performance of the three-class algorithm; vertical columns add to 100%, to indicate the proportion of predicted measurements that were placed into the correct category. The tree-based classifier performed very well in distinguishing the non-neoplastic cases (79.3%) and when predicting HGD and EAC (70.4% and 90.3%, respectively).

### 4 Discussion

In this study, we demonstrated the use of a simple, low-cost, portable, high-resolution microendoscopy system to distinguish between clinically significant grade of Barrett’s esophagus using subjective visual interpretation and objective quantitative image analysis and classification. Subjective analysis of the images by expert gastroenterologists achieved average sensitivity and specificity of 87% and 53%, respectively, while expert pathologists achieved average sensitivity and specificity of 87% and 68%, respectively. Image classification algorithms were created by analysis of key textural features within the images, the most important features being frequency content, as calculated by a discrete fast Fourier transform, and pixel pair correlation, as calculated by a gray-level co-occurrence matrix. The objective classification algorithm was able to distinguish between neoplastic and non-neoplastic cases with a sensitivity of 87% and a specificity of 85%.

The results of this pilot study suggest that this technique may be useful to regions without highly trained expert personnel or extensive biomedical infrastructure. The low cost of the device (roughly US $2,500) and the wide availability of low-cost computers may facilitate the distribution of the HRME and the quantitative image analysis and classification algorithm to regions beyond tertiary care centers and university hospitals.

### Table 3 Summary of the performance of the tree-based three-class classifier.

<table>
<thead>
<tr>
<th></th>
<th>IM/LGD by histology</th>
<th>HGD by histology</th>
<th>EAC by histology</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM/LGD predicted</td>
<td>79.3%</td>
<td>7.4%</td>
<td>2.3%</td>
</tr>
<tr>
<td>HGD predicted</td>
<td>8.6%</td>
<td>70.4%</td>
<td>4.7%</td>
</tr>
<tr>
<td>EAC predicted</td>
<td>12.1%</td>
<td>22.2%</td>
<td>93.0%</td>
</tr>
</tbody>
</table>
Subjective scoring of the HRME images identified that significant intra-observer variability exists. While the overall accuracy is comparable to the results achieved by the computational diagnostic algorithm presented here, the subjective interpretation result demonstrates that even among highly trained endoscopists (familiar with microendoscopy) and expert gastrointestinal (GI) pathologists, disagreement can occur. Unfamiliarity with the image modality presented here by the HRME may be a confounding factor, as can bias from extensive experience with more conventional modalities, such as histopathology. The objective algorithm presented here may be able to aid clinicians in making diagnostic decisions at the point of care, while reducing intra-observer variability.

As with any small pilot study, a larger sample size will be required to optimize and evaluate the performance of the classification algorithm presented here. Results with the relatively small number of patients (9) and measurement sites (128), as well as the distribution of diagnoses measured (Table 2) may not reflect the HRME probe’s performance in a screening setting. The non-neoplastic sites examined here included both intestinal metaplasia and low-grade dysplasia; however, the number of sites with low-grade dysplasia (53) was much greater than the number with intestinal metaplasia (6). While this is expected in the high-risk population participating in this study, it may induce bias in the accuracy rates reported in this feasibility study. In a normal screening population, the proportion of non-neoplastic sites with intestinal metaplasia only is expected to be higher than that encountered here. We are evaluating the performance of the algorithm in a larger number of patients in a normal screening population of average risk. A separate, independent validation data set is the most robust method to verify the performance of the HRME device and classification algorithm. However, in the absence of an independent validation set, cross-validation is recognized as a helpful tool to guard against overtraining when estimating the performance of a classifier.

An additional limitation of this trial is the HRME probe’s inability to image subsurface regions. While the incidence of high-risk submucosal or occult EAC is still subject to controversy, inspection of subsurface regions remains an important consideration to clinicians. The current HRME device yields images by placing the tip of the device into direct contact with tissue, limiting its image acquisition to the superficial epithelium. However, the fiber-optic probe is small enough to pass through the lumen of a 16-gauge needle; the needle could be used to physically penetrate into deeper layers of the epithelium, allowing the fiber bundle to image these areas. This technique has been successfully demonstrated in a mouse model.

The quantitative algorithm described in this paper could be applied to any high-resolution imaging system capable of digital image acquisition. In particular, confocal imaging, which has shown much promise as an in situ high-resolution imaging device, could benefit from objective image analysis. While the same features that were demonstrated to be diagnostic for this study (frequency content and pixel pair correlation) may not be applicable to all high-resolution imaging systems, a similar textural feature search could employed. The new algorithm may then be used to augment physician diagnosis during a procedure.

High-resolution imaging of the gastrointestinal epithelium offers clinicians a means to inspect the histologic features of suspicious lesions and post-resection margins in real time during an endoscopic procedure. Such information can be used to guide the selection of sites for biopsy to improve the diagnostic yield during screening endoscopy. Traditional confocal laser endoscopy (CLE) typically requires the endoscopist to interpret images during the procedure. This subjective interpretation is likely highly dependent on the endoscopist’s decision as to the classification of a lesion. Such objective image analysis has been applied in other high-resolution imaging studies; Becker et al. demonstrated that computer-aided diagnosis of mucosal pit patterns but a fiber bundle confocal system has been shown to be effective in distinguishing neoplastic from non-neoplastic Barrett’s tissue.

While the small field of view (750 microns) of the HRME device makes surveillance of the entire esophageal mucosa impractical, the HRME could be used as a tool to investigate suspicious sites during an endoscopic procedure. Standard white-light endoscopy has a field of view of several centimeters and can be used to identify regions of interest that require high-resolution interrogation with the HRME probe. In addition, other wide-field imaging techniques are becoming available, such as narrow-band imaging (NBI) and autofluorescence (AF) imaging, which could be useful in uncovering occult premalignant conditions. Such techniques are susceptible to confounding factors, such as inflammation, which can be misclassified as dysplasia. High-resolution imaging of these regions may be able to correctly differentiate benign inflammatory processes from precancerous lesions, resulting in improved biopsy localization and higher diagnostic yield.

The image processing techniques presented in this study can uncover features in images that are clinically important, but that may not be apparent to the clinician performing the procedure. In addition, quantitative image analysis and classification techniques require relatively minimal computing time, which is useful in a busy clinical setting. This feasibility study demonstrates that images acquired with a low-cost, high-resolution microendoscope can provide objective discrimination of neoplastic and non-neoplastic sites in the esophagus with good accuracy relative to histology. Prospective evaluation of this approach is warranted to determine the clinical performance in a low-risk screening setting.

Acknowledgments

The authors would like to acknowledge funding from NIH Grant Nos. 1RO1EB007594-01 and 1RO1CA103830-05. One of the co-authors (R.K.) has a small ownership interest in Remicalm, Inc., which has licensed related technology from the University of Texas at Austin and Rice University.

References


