SPECTRAL CYTOPATHOLOGY: A DIAGNOSTIC TOOL FOR THE ORAL CAVITY AND THE ESOPHAGUS

A thesis presented by

Kathleen Lenau

to
The Department of Chemistry and Chemical Biology

In partial fulfillment of the requirements for the degree of
Master of Science

in the field of
Chemistry

Northeastern University
Boston, Massachusetts
May 3, 2013
SPECTRAL CYTOPATHOLOGY: A DIAGNOSTIC TOOL FOR THE ORAL CAVITY AND THE ESOPHAGUS

A thesis presented
by

Kathleen Lenau

ABSTRACT OF THESIS

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Chemistry in the College of Science of Northeastern University
May 3, 2013
ABSTRACT

Spectral cytopathology (SCP), which couples Fourier Transform infrared microspectroscopy (IR-MSP) with unsupervised multivariate statistical methods, is an accurate and robust optical tool for the detection of subtle biochemical changes that occur within individual cells. SCP monitors the biochemical composition of cells within a sample and through computer-based algorithms is able to identify and classify unique spectral changes that are induced by disease. These subtle molecular changes can be observed before morphological changes become apparent, allowing earlier diagnostic information to become available for cytologists. When diagnostic information is available earlier the patient can potentially be offered treatment more swiftly, thus improving the prognosis for the patient.

In this thesis, SCP was used for two different applications; 1) investigation of influential factors for the oral cavity and 2) diagnosis of Barrett’s esophagus. For both projects infrared spectra were collected from a 4 mm x 4 mm area of cells that had been deposited onto MirrIR slides. Spectroscopic data were processed using the “SkinnyCells” MATLAB program developed in-house combining previous data processing techniques into one automated method. Spectral data were analyzed using unsupervised multivariate methods, specifically Principal Component Analysis (PCA) to determine disease characteristics, and a supervised training algorithm to provide a diagnosis of normal or abnormal.

The first project began in 2010 when a pre-clinical trial was initiated at Northeastern University to further investigate applications of SCP for the oral cavity. The goals of the pre-clinical trial were to expand previous databases, and develop a pre-screening tool for oral disease. In total, 93 volunteers contributed their oral squamous cells resulting in 279 samples in addition to clinical samples provided by Tufts Medical Center. Results from the pre-clinical trial
showed that SCP was able to distinguish spectral changes related to anatomical region, viral infection, oral disease, and medications.

For the second project collaboration with Department of Oesophagogastric Surgery at the Gloucestershire Hospitals NHS Trust provided samples from the esophagus to test SCP as a diagnostic tool for Barrett’s esophagus. Barrett’s esophagus is known as the major risk factor for the development of esophageal adenocarcinoma, for which prognosis is very poor. SCP was used to distinguish between two cell types: normal squamous and those collected from an area of Barrett’s esophagus. Through the application of SCP it is hoped that cytologists would be able to diagnose a patient with Barrett’s esophagus earlier using easier and faster methods to allow treatment to control the disease before it progresses to esophageal adenocarcinoma.
ACKNOWLEDGEMENTS

First, and foremost, I would like to thank my family and friends for all of their support, encouragement and guidance throughout my five years at Northeastern University. I am sincerely grateful to my parents, Beth and Bob, who have given me all of the necessary tools to be successful during my many years of school. I am so grateful for their love and support in everything I have set my mind on doing in my life, especially when I chose to attend a University that was 3,000 miles from my home on the West Coast. I would also like to extend thanks to all of my Aunts and Uncles on the East Coast who took me in on multiple weekends when I had to get away from the city.

I owe special thanks to my research advisor, Dr. Max Diem, who gave me the chance as a sophomore to come into his laboratory and be a member of his research group. It has been a privilege to work in this laboratory over the past four years and to have learned so much during this time.

I would like to thank every member of the Laboratory for Spectral Diagnosis at Northeastern University, who have all helped me so much during my time in the lab. Thank you to Max Diem, Jennifer Schubert, Jennifer Fore, Miloš Miljković, Benjamin Bird, Antonella Mazur, Christina Uttero, Ellen Marcislin, Evgenia Zuser, Tatyana Chernenko, and Doug Townsend. I would like to acknowledge all of the help I received from our collaborators on the esophageal project at the Department of Oesophagogastric Surgery in the Gloucestershire Hospitals NHS Trust, specifically Dr. Max Almond with whom I was in constant contact, and always answered my emails so quickly.

Special thanks go to my thesis committee: Dr. Max Diem, Dr. Max Almond, and Dr. Patricia Mabrouk.
I want to thank the Department of Chemistry and Chemical Biology at Northeastern University, for the education, opportunities, and support provided to me over the years. Thank you to Kathleen Cameron, Alex Henriksen, John Bottomy, Graham Jones, Rich Pumphrey, and Cara Shockley. I would also like to specifically thank Jordan Swift, who as my original academic and co-op advisor, was always there for me when I had questions or concerns about anything to do with classes or my career, and also for being a wonderful friend to me as well.
# TABLE OF CONTENTS

ABSTRACT OF THESIS .................................................................................................................. 2

ACKNOWLEDGEMENTS ................................................................................................................. 5

TABLE OF CONTENTS ................................................................................................................... 7

LIST OF FIGURES .......................................................................................................................... 10

LIST OF ABBREVIATIONS .............................................................................................................. 12

CHAPTER 1: INTRODUCTION TO SPECTRAL CYTOPATHOLOGY ............................................. 14

  1.1 Current Methods for Disease Diagnosis .................................................................................. 14
  1.2 Spectral Cytopathology .......................................................................................................... 14
  1.3 References .............................................................................................................................. 16

CHAPTER 2: EXPERIMENTAL METHODS ..................................................................................... 18

  2.1 Sample Collection .................................................................................................................... 18
  2.1.1 Oral Cavity Samples ......................................................................................................... 18
  2.1.2 Barrett’s Samples ............................................................................................................. 21
  2.2 Sample Preparation ................................................................................................................ 21
  2.3 Data Acquisition ..................................................................................................................... 24
  2.4 Data Processing ...................................................................................................................... 26
  2.4.1 Noise-Adjusted Principal Component Reconstruction .................................................... 27
  2.4.2 Phase Correction ............................................................................................................. 29
  2.4.3 Water Vapor Correction .................................................................................................. 31
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1.1</td>
<td>The Disease</td>
<td>56</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Current Diagnostic Methods</td>
<td>58</td>
</tr>
<tr>
<td>4.1.3</td>
<td>Current Treatment Methods</td>
<td>60</td>
</tr>
<tr>
<td>4.2</td>
<td>Preliminary Results</td>
<td>61</td>
</tr>
<tr>
<td>4.3</td>
<td>Future Directions</td>
<td>66</td>
</tr>
<tr>
<td>4.4</td>
<td>Conclusions</td>
<td>66</td>
</tr>
<tr>
<td>4.5</td>
<td>References</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>APPENDIX A: REPRINT PERMISSIONS FOR FIGURES</td>
<td>70</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 1.1  Spectra of Biochemical Components of Cells
Figure 2.1  SurePath Fixative Vial and Cytobrush
Figure 2.2  Oral Pre-Clinical Trial Questionnaire
Figure 2.3  Cytospin Instrumentation and Conical Funnel
Figure 2.4  MirrIR Substrates
Figure 2.5  Sample Spot on MirrIR Substrate
Figure 2.6  Perkin Elmer Spectrum One/Spotlight 400 Infrared Micro-Spectrometer
Figure 2.7  Visual Image of Sample Area
Figure 2.8  Visual Image of One Sample Quadrant, IR Intensity Map, and Raw Cellular Spectrum
Figure 2.9  Noise-Adjusted Principal Component Reconstruction of Cellular Sample
Figure 2.10 Mie and R-Mie Scattering Effects
Figure 2.11 Visual Image, IR Absorbance, and Binary Mask in PapMap
Figure 2.12 Application of Filter Program
Figure 2.13 Example of 2^nd Derivative Application on Absorbance Data
Figure 2.14 Principal Component Analysis Scores Plots
Figure 2.15 Artificial Neural Network Design
Figure 2.16 Pap Stained Cellular Sample and Visual Microscope
Figure 3.1  PCA Scores Plot for Clinical
Figure 3.2  First Principal Components
Figure 3.3  ANN Results for the Tongue
Figure 3.4  PCA Scores Plot for Smoker
Figure 3.5  PCA Scores Plot for Oral Contraceptive
Figure 3.6  PCA Scores Plot for Aspirin
Figure 3.7  PCA Scores Plot for Mononucleosis
Figure 4.1  Transformation of Esophageal Lining due to Barrett’s Esophagus
Table 4.1  Risk Factors for Barrett’s Esophagus
Figure 4.2  PCA Scores Plots
Figure 4.3  Mean Spectra for Normal and Barrett’s Esophagus
Figure 4.4  Artificial Neural Network Results
## LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>BE</td>
<td>Barrett’s Esophagus</td>
</tr>
<tr>
<td>BSS</td>
<td>Hank’s Balanced Salt Solution</td>
</tr>
<tr>
<td>EAC</td>
<td>Esophageal Adenocarcinoma</td>
</tr>
<tr>
<td>EMR</td>
<td>Endoscopic Mucosal Resection</td>
</tr>
<tr>
<td>EMSC</td>
<td>Extended Multiplicative Signal Correction</td>
</tr>
<tr>
<td>EPSS</td>
<td>Endoscopic Polarized Scanning Spectroscopy</td>
</tr>
<tr>
<td>FT</td>
<td>Fourier Transform</td>
</tr>
<tr>
<td>FT-IR</td>
<td>Fourier Transform Infrared Spectroscopy</td>
</tr>
<tr>
<td>HGD</td>
<td>High-grade Dysplasia</td>
</tr>
<tr>
<td>HPV</td>
<td>Human <em>Papillomavirus</em></td>
</tr>
<tr>
<td>IR</td>
<td>Infrared</td>
</tr>
<tr>
<td>IR-MSP</td>
<td>Infrared Micro-Spectroscopy</td>
</tr>
<tr>
<td>IRB</td>
<td>Institutional Review Board</td>
</tr>
<tr>
<td>LGD</td>
<td>Low-grade Dysplasia</td>
</tr>
<tr>
<td>Low-e</td>
<td>Low-emissivity</td>
</tr>
<tr>
<td>NA-PC</td>
<td>Noise-Adjusted Principal Component Reconstruction</td>
</tr>
<tr>
<td>NSAIDS</td>
<td>Non-Steroidal Anti-Inflammatory Drugs</td>
</tr>
<tr>
<td>PC</td>
<td>Principal Component</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PE400</td>
<td>Perkin Elmer Spectrum One/Spotlight 400 Imaging IR Micro-Spectrometer</td>
</tr>
<tr>
<td>PPI</td>
<td>Proton Pump Inhibitor</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>R-Mie</td>
<td>Resonant Mie Scattering</td>
</tr>
<tr>
<td>S/N</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SCP</td>
<td>Spectral Cytopathology</td>
</tr>
</tbody>
</table>
CHAPTER 1: INTRODUCTION TO SPECTRAL CYTOPATHOLOGY

1.1 Current Methods for Disease Diagnosis

Current methods utilized in hospitals for the diagnosis of various diseases oftentimes cannot provide accurate diagnoses until late stages, and may also provide false results. The typical screening method employed is known as cytopathology and involves a visual inspection of a cellular sample. This generates inherent flaws because of the reliance on an individual to visually identify small morphological differences in the nucleus to cytoplasm ratio, size and shape of the nuclear membrane and chromosomal granularity within a sample containing tens of thousands of cells, where approximately 10 of which are diagnostically relevant. In order to assist the cytologist in making an appropriate diagnosis, staining procedures are often used, such as the Pap stain. These protocols stain various cellular structures to aid the cytologist in diagnosing the clinical samples. However, a cytologist may still misdiagnose the sample due to the presence of a few diseased-looking cells contained in the mostly normal sample. [1]

1.2 Spectral Cytopathology

FTIR microspectroscopy (IR-MSP) has been shown to be a less subjective, automatable, label-free, and non-destructive method for fast, and easy disease diagnosis. Utilizing unsupervised multivariate statistical methods of analysis in conjunction with IR-MSP allows for the detection of small biochemical changes within an entire sample without having to rely on the morphological changes that conventionally indicate disease. [2-10]

IR-MSP measures the vibrations that occur within a molecule, caused by the absorption of a photon of light at a specific frequency which is equal to the vibrational frequency of a bond
within that molecule. IR-MSP generates an infrared spectrum using a Fourier transform, which allows for the whole wavelength range to be sampled at once, to convert the raw data into a spectrum. An infrared (IR) spectrum is produced when a molecule absorbs an infrared photon and becomes excited into higher vibrational states. In order for a molecule to absorb IR light it must undergo a net change in dipole moment, which occurs when there is a variation between the centers of positive and negative charges within the molecule. [3,4]

Spectral cytopathology (SCP) is a term that refers to the use of IR-MSP coupled with unsupervised multivariate statistical methods of analysis as a means of disease diagnosis. SCP uses IR-MSP to detect the spectral patterns that are related to molecular changes within cells characteristic of disease. This technique could allow for earlier diagnosis and treatment due to its ability to detect the biochemical changes associated with disease before the disease manifests morphologically. It is a common misconception that IR-MSP is able to distinguish between the different biochemical characteristics of a cell, while in actuality a superposition of the spectra for all biochemical components is observed. For example, IR-MSP can detect changes in proteins overall within a cell but cannot differentiate between specific proteins. A few common cellular components are shown spectrally in Figure 1.1, along with their representative infrared spectra. While SCP has been shown to be an accurate and effective diagnostic tool in the Laboratory for Spectral Diagnosis [4-8], it is still necessary to couple the method with conventional cytopathology in order to properly correlate results.
Figure 1.1 (A) The spectra linked to four different biochemical components of a cell, with (B) the observed superposition of these components. Figure courtesy of Dr. Max Diem.

1.3 References


CHAPTER 2: EXPERIMENTAL METHODS

2.1 Sample Collection

2.1.1 Oral Cavity Samples

Clinical samples from the oral cavity were obtained through collaboration with Tufts Medical Center in the Department of Pathology and Otolaryngology [Boston, MA, USA]. Tufts Medical Center provided a demographic sheet for each sample with details such as the dates of sample collection, the patient’s clinical history and surgical pathology reports. The clinical samples were collected by swabbing the site of interest with a cytobrush and then immersing the brush in SurePath fixative solution [Burlington, NC, USA], containing 24% aqueous ethanol and 1% each of methanol and 2-propanol. The SurePath solution is a fixative, which preserves the sample by removing lipids and making proteins insoluble through dehydration. [1-2] An example of the SurePath fixative solution and cytobrush is provided in Figure 2.1.

Figure 2.1 Vial containing SurePath fixative solution with accompanying cytobrushes used to collect and store cellular specimens both at Tufts Medical Center and the Laboratory for Spectral Diagnosis.
In 2010 the Laboratory for Spectral Diagnosis began a pre-clinical trial focusing on the oral cavity with the goal of expanding the database for normal oral cavity samples. Before the samples were collected, the Laboratory for Spectral Diagnosis obtained approval from the Institutional Review Board (IRB) at Northeastern University.

For the pre-clinical trial, samples were collected from three regions of the oral cavity; the cheek, tongue, and mouth floor. In all, 93 volunteers from Northeastern University’s campus donated their oral squamous cells for each of the three regions yielding a total of 279 samples. Along with obtaining the samples, volunteers were asked to fill out an informational questionnaire, shown in Figure 2.2, that detailed information such as gender, race, age, tobacco usage, medication usage, history of oral disease, mononucleosis, *herpes simplex* virus, and if they had been vaccinated against HPV with Gardisil. To keep volunteer information private, barcodes were placed on each questionnaire, sample vial, and prepared slide.

Once the volunteers had given their consent and filled out the questionnaire they rinsed their mouths with water to remove any food and debris. Next, each volunteer was provided with one cytobrush per region to be swabbed. Volunteers gently swabbed each area and placed the cytobrush into individual vials containing SurePath solution.
Oral Cytology Research Questionnaire

Northeastern University (NU), in collaboration with Tufts Medical Center (Boston, MA), is developing new ways to screen oral cell samples for abnormalities, such as pre-cancer and viral infections. This study was funded by a grant from the Cancer Institute of the National Institutes of Health (NIH). We would like to ask you to participate in this study, which is for research purposes only. For this study, a sample will be collected by swabbing the inside of your mouth, specifically the inside of your cheek, the surface of your tongue, and under your tongue, with a cytological brush. In addition, we ask you to answer a few questions. Your participation is voluntary, and your identity and the results from your participation will remain anonymous and confidential.

<table>
<thead>
<tr>
<th>Gender (circle)</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race (circle)</td>
<td>Caucasian</td>
<td>African</td>
</tr>
<tr>
<td>Age:____________</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do you smoke:</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>How long since the last cigarette (minutes/hours)?______________</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do you chew tobacco?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Have you taken any over the counter or prescription medication during the last 24 hours?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>What medication?______________________________</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do you presently have an active cold sore?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Have you recently had a cold sore?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Have you had mononucleosis (“mono”) in the past 5 years?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Have you had any history of abnormal oral infections? (e.g. herpes or HPV infection, cancer)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Have you received Gardasil (HPV vaccine)?________</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

**Figure 2.2** Informational questionnaire given to volunteers for the oral pre-clinical trial.
2.1.2 Barrett’s Samples

Collaboration with the Department of Oesophagogastric Surgery at the Gloucestershire Hospitals NHS Trust [Gloucestershire, UK] was established in order to investigate the use of SCP as an early diagnostic tool for Barrett’s esophagus. Ethical approval for the study was granted by the Gloucestershire Local Research Committee, [Gloucestershire, UK], and samples were collected during endoscopy after written consent was obtained from each patient. An endoscopic cytology brush was used to collect the patient’s cells by passing it through the endoscope and brushing against the desired site of the esophagus. Immediately following exfoliation, the cytobrush containing the cells was immersed in formalin. Formalin is another fixative solution that cross-links primary amino groups within the protein via an adjacent nitrogen atom. [3] The sample vials containing the formalin fixed cells were then shipped to the Laboratory for Spectral Diagnosis for processing and data analysis. In total, 23 samples from 11 patients were received; 8 of which were diagnosed as Barrett’s esophagus, 2 with Barrett’s esophagus and possible dysplasia, 11 normal squamous, 1 from the stomach, and 1 with esophageal adenocarcinoma. However, cells from the samples collected from the stomach and those with esophageal adenocarcinoma were unable to be salvaged due to low cell counts.

2.2 Sample Preparation

Regardless of the fixation method utilized, be it SurePath or formalin, samples remained in the fixative for at least a week before sample preparation began to mimic hospital conditions. In order to best remove the cells from the cytobrush bristles, the vials were vortexed for approximately 30 seconds followed by removal of the cytobrush. The solution containing the cells was then transferred to a 15 mL centrifuge tube and diluted to 5 mL with the respective
fixative solution. The tube was centrifuged for 15 minutes to produce a cell pellet at the bottom of the tube. Once centrifugation was complete, the supernatant was removed with a glass pipette in a manner so as not to agitate the cell pellet. The cell pellet was re-suspended in 5 mL of Hank’s Balanced Salt Solution (BSS) 1X [ATCC, Manassas, VA, USA] and vortexed for 30 seconds. The washing process was completed twice with BSS. After the final centrifugation step, the cell pellet was diluted with 2 mL of BSS.

A conical funnel was cleaned with dry, compressed air to remove any excess dust or debris and placed into a Cytospin [Thermo, Waltham, MA, USA], shown in Figure 2.3. Cytocentrifugation was performed at 800 rpm for 12 minutes in order to deposit the cells onto low emissivity (low-e) MirrIR microscope slides [Kevley Technologies, Chesterland, OH, USA]. MirrIR slides, shown in Figure 2.4, were chosen because they are transparent in visible light and reflective in the IR, while being relatively inexpensive, at approximately $2.00 per slide. [4] The cytocentrifugation method applies centrifugal force to the solution within the conical funnel in order to deposit cells onto the MirrIR slide while the excess liquid is absorbed by the paper surrounding the funnel. The conical funnels contain a small circular opening which allows for a monolayer of cells to be deposited in a 5 mm diameter onto the MirrIR slides, exhibited in Figure 2.5. The cellularity of the sample solution determined the amount deposited into the conical funnel. The slides were allowed to dry and the cells to adhere for 10 minutes post cytocentrifugation. The slides were then submerged in Millipore water 5 times to remove any salt crystals that may have formed due to the presence of BSS and allowed to dry for an additional 10 minutes. Prepared slides were stored in a desiccator for a minimum of 24 hours before data acquisition in order to ensure that excess water vapor would not be present to negatively affect the acquired spectra.
Figure 2.3 (A) Cytospin instrument into which (B) plastic conical funnels are placed in order to deposit cells onto the surface of MirrIR slides.

Figure 2.4 MirrIR microscope slides utilized as sample substrates due to their low cost and reflectivity/transparency properties.

Figure 2.5 Cells deposited in a 5 mm diameter onto the MirrIR slide.
2.3 Data Acquisition

A Perkin Elmer Spectrum One/Spotlight 400 (PE 400) imaging IR micro-spectrometer, shown in Figure 2.6, was used to collect data in imaging mode from a 4 mm x 4 mm area of the sample spot averaging 1,000 cells. In imaging mode, the PE 400 utilized a 16-element focal plane array detector. Within the sample spot, four 2 mm x 2 mm quadrants were designated and the accompanying visual image of each quadrant was collected, shown in Figure 2.7. The sample spot was raster scanned at a pixel size of 6.25 μm to obtain an IR image of each quadrant with the accompanying absorbance data, exhibited in Figure 2.8. The instrument was operated in transflection mode at a 4 cm\(^{-1}\) spectral resolution, with the co-addition of 2 interferograms per pixel in the spectral range of 4000-750 cm\(^{-1}\). The background was collected at 120 scans from a position where no cellular material was present. The acquired data were Fast Fourier Transformed using a Norton-Beer apodization [5] with one level zero-filling. Data acquisition required approximately 6-8 hours to collect images for all four of the 2 mm x 2 mm quadrants within a sample. Raw datasets in the form of *.fsm files which were 635.8 MB were then transferred from the instrument to a computer workstation for data analysis.

Figure 2.6 The Perkin Elmer Spectrum One/Spotlight 400 imaging IR micro-spectrometer used
in the Laboratory for Spectral Diagnosis.

Figure 2.7 Visual image of the 4 mm x 4 mm image chosen for data analysis divided into 4 equal quadrants, with quadrant 1 situated at the top left, quadrant 2 at the top right, quadrant 3 at the bottom left, and quadrant 4 at the bottom right.
Figure 2.8 (A) Visual image of a sample with quadrant 3 selected for image mode data collection. (B) The infrared absorption intensity color map from quadrant 3 in A, and (C) an example of the raw infrared absorbance spectrum for a cell in the sample.

2.4 Data Processing

A fast method of data processing, entitled “SkinnyCells” was created by Dr. Miloš Miljković within the Laboratory for Spectral Diagnosis utilizing MATLAB [Mathworks, Natick, MA, USA]. This technique combined the previously utilized data processing programs within
MATLAB into one convenient, automated, and user-friendly procedure. Previously, it required roughly one hour to perform all data processing routines, but “SkinnyCells” has now decreased that time to approximately two minutes. This section breaks down each portion of the “SkinnyCells” program and how they were each utilized together to produce superior spectra.

2.4.1 Noise-Adjusted Principal Component Reconstruction

The first portion of the “SkinnyCells” program improves the spectral signal-to-noise (S/N) ratio of the collected infrared spectra. Noise-Adjusted Principal Component Reconstruction (NA-PC) is a MATLAB based algorithm developed in the Laboratory for Spectral Diagnosis based on previous literature. [6-7] The Laboratory for Spectral Diagnosis has demonstrated NA-PC as a reliable method to improve the spectral S/N ratio of data. [6-8] The first step of NA-PC was to apply a baseline-offset to the spectra, so the lowest intensity point in the spectral range of 2200-1480 cm\(^{-1}\) for the oral data, and 2200-1900 cm\(^{-1}\) for the Barrett’s esophagus data, was set to zero as the baseline. These regions were used as they are where typical protein vibrations are observed in an IR spectrum. Next, a quality test was performed on the baseline-offset spectra in order to generate an intensity map using the spectral region of 1676-1626 cm\(^{-1}\), and a “noise co-variance matrix” was created to represent solely the noise within the sample. To create the “signal co-variance matrix”, which would represent exclusively the “signal” within the sample, the original spectra were again baseline-offset corrected. A quality test was then performed in the 1676-1626 cm\(^{-1}\) spectral region, and this time the intensity map was used to determine where biological material was present within the sample. Another quality test in the entire spectral range was performed to eliminate spectra that could be considered “cellular debris”.

27
Eigenvector matrices were computed from each of the “noise” and “signal” co-variance matrices. The eigenvectors for the “signal co-variance matrix” were then realigned to place the noise eigenvectors at the end. Reconstruction of the individual pixel spectra used the first 30 principal components. [9] Figure 2.9 shows the individual steps of NA-PC for a cellular specimen.

Figure 2.9 (A) Noise and (B) Signal sample areas for the calculation of NA-PC to enhance the spectral S/N ratio. Also shown are the spectra (C) before and (D) after NA-PC reconstruction.
2.4.2 Phase Correction

Spectra collected via IR-MSP are often subjected to various scattering effects, which have been noted with increasing incidence over the past several years. [10-12] These scattering effects include Mie scattering, reflection contributions, and Resonant Mie (R-Mie) scattering, and each effect causes distortions in the collected spectra that have an adverse effect during data analysis. [10, 13-14] Much effort has been exerted in order to understand and correct for these scattering effects. [10, 13-17]

Mie scattering is observed when the shape and size of individual cells and their nuclei are similar to the wavelength of IR radiation. Mie scattering causes a broad sinusoidal oscillation that appears in the baseline of the spectra, causing a misrepresentation in the position and intensity of absorption bands. [13-14] R-Mie and reflection scattering contributions both affect the band shape on the high wavenumber side of the amide I band, cause a peak shift of the amide I band, and an abnormal peak height ratio between the amide I and amide II bands. [12-13] These contributions are exhibited in Figure 2.10.
Figure 2.10 Example of scattering effects observed in absorbance spectra caused by Mie and R-Mie scattering as compared to a normal absorbance spectrum. Figure courtesy of Dr. Max Diem. [15]

The Laboratory for Spectral Diagnosis has utilized many methods in the past in order to correct for these three scattering contributions. Currently, the method of correction used is known as “phase correction” [16] and is currently pending U.S. patent protection [17]. Phase correction was performed on spectra having 512 data points that had been converted to 2nd derivative, which were then reverse Fourier transformed to obtain interferograms. The real and imaginary parts of the interferograms were then shifted by a phase angle thereby removing the R-Mie based scattering effects.
2.4.3 Water Vapor Correction

The instruments used in the laboratory, although constantly purged with dry air, may still be contaminated by water vapor. Water vapor contributions interfere with absorbance spectra of cellular samples in the region of 1800-1350 cm\(^{-1}\), where most protein vibrations are observed. A more preferable method than the algorithm provided in the PE 400 to correct for water vapor contributions was adapted from literature [18], and was developed in the Laboratory for Spectral Diagnosis.

In the water vapor correction portion of the “SkinnyCells” program, extended multiplicative signal correction (EMSC) was applied to the spectra in order to remove the contribution seen in the spectra from atmospheric water. This created an interference spectrum from the higher wavenumber region of the spectra to model the water vapor by averaging a few interferograms. The interference spectrum was then used to deduct any atmospheric contributions from the spectra.

2.4.4 PapMap

To assist in the speed of data acquisition and processing, a MATLAB based program, known as PapMap, was developed in-house. [19-20] PapMap is used to reconstruct the spectra of individual cells by calculating the cellular spectrum for each area that a cell occupies beginning with the spectrum that has the largest amide I intensity. PapMap works to recognize which pixels correspond to a particular cell and averages those pixels following certain guidelines in order to produce one spectrum per cell.

To determine which pixels belong to a cell, a binary mask was used, which is exhibited in
Figure 2.1. A selected number of pixels, typically between 10 and 100, was set to equal the size of a cell. Anything larger or smaller than this set pixel size was considered to be either a clump of cells or debris and was discarded. Next, spectra that constituted a cell were averaged after elimination of pixels of low S/N and edge effects. This was done by adding a restriction to regulate peak location and peak intensity, (set to 1670-1630 cm\(^{-1}\)). The spectrum with the highest amide I intensity typically does not include any R-Mie or edge effects, and was set as the “seed spectrum”. If spectra were above 67% for the amide I intensity and the frequency was within a wavenumber shift of 4 cm\(^{-1}\) of this “seed spectrum”, then that spectrum was included in the process of averaging together the spectra for each cell.

![Figure 2.1](image)

**Figure 2.11** (A) Visual images of cells collected on the PE 400 with (B) the color map indicating amide I intensity, and (C) the accompanying binary mask created in PapMap.

### 2.4.5 Spectrum Filter

Although there were many other methods applied to remove poor quality spectra within “SkinnyCells”, some were not caught by these quality controls. An example of a dataset contaminated by these spectra, particularly straight lines, is shown in Figure 2.12. It was still important to remove any remaining poor quality spectra so that they did not adversely affect the results by falsely contributing to the differentiation between classes determined by Principal
Component Analysis (PCA). This could be done by hand but requires a large amount of time to look through each spectrum individually to decide which did not contain pertinent cellular information. In order to reduce the time required and the subjectivity from human involvement, a MATLAB function was created by Dr. Antonella Mazur. The “filter” function works by taking a vector file containing equal numbers of cellular spectra from each class being investigated. A comparison between the vector file and the spectra investigated was performed by calculating a correlation coefficient in the spectral region of 3540-2840 cm\(^{-1}\). If the correlation coefficient was calculated to be greater than 85%, the spectra were retained and considered to have valuable cellular information. Those which had a correlation coefficient of less than 85% were removed from the overall dataset.

![Figure 2.12](image.png)

**Figure 2.12** (A) Example of a dataset contaminated with poor spectra, such as straight lines, and (B) the resulting dataset after removal of the poor spectra using the “filter” function in MATLAB.

### 2.4.6 Smoothing Derivative Filter

Rather than looking at individual peaks within a spectrum a superposition of all biochemical components of a cell are observed for the collected spectra. The slight changes in
the spectra between normal and abnormal classes may be difficult to observe and interpret due to this overlapping. In data processing, spectra were converted to 2\textsuperscript{nd} derivative to enhance these differences to allow for easier interpretation. The differences were enhanced because when a spectrum is converted into 2\textsuperscript{nd} derivative the bandwidth collapses, enhancing spectral resolution, and allowing for previously unnoticed peaks to emerge. The conversion to 2\textsuperscript{nd} derivative also removed any sloping baseline features (Figure 2.13).

The spectra were also subjected to smoothing using a Savitzky-Golay smoothing window of 13 points to assist with reducing spectral noise. [21] Finally, the spectra were vector normalized in order to remove any dependence on differences in sample thickness and density.

![Figure 2.13](image)

Figure 2.13 (A) Representation of an absorbance spectrum (top) which was converted to 2\textsuperscript{nd} derivative (bottom). (B) Conversion from absorbance (left) to 2\textsuperscript{nd} derivative (right) showing the enhancement of the subtle differences that may not be apparent in absorbance spectra. Images provided by Dr. Max Diem.
2.5 Data Analysis

2.5.1 Principal Component Analysis

The primary unsupervised multivariate method of data analysis utilized by the Laboratory for Spectral Diagnosis for cellular samples is Principal Component Analysis (PCA) and was performed in MATLAB using the PLS ToolBox 6.0 [Eigenvector Research, Wenatchee, WA, USA]. PCA is a dimension reduction technique, where points in multidimensional space are projected onto a space of fewer dimensions. The goal of PCA is to reconstruct the spectra in terms of their eigenvectors (also referred to as principal components (PCs), or loading vectors), then plot the degree of their contribution in a “scores plot”, as shown in Figure 2.14. The first PC, which is extracted from the dataset, contains the largest variance, the second PC contains the second largest variance, and so on, with the majority of the variance enclosed within the first few PCs. It is also possible to view the individual PCs to observe where the variance between datasets lies. PCA is useful because much of the variance that is detected by the unsupervised multivariate method cannot be detected through viewing the spectra with the naked eye. When the dataset is reconstructed based on the combination of the PCs, the scores indicate the contribution of each PC to the reconstructed spectra. If there are no spectral differences detected by PCA, then the scores plot appears to be a “shotgun blast”, where the data points are scattered with no apparent pattern. If there are spectral differences detected, then the data points will be split and a pattern of separation will emerge.
Figure 2.14 PCA scores plots for a sample exhibiting (A) spectral variance and (B) random variance, known as a “shotgun blast”.

2.5.2 Artificial Neural Networks

Diagnostic algorithms, such as Artificial Neural Networks (ANN), have been used as supervised statistical methods previously in the Laboratory for Spectral Diagnosis for the analysis of IR spectral data sets for tissue specimens. [9,22] ANNs are modeled after the neural patterns found in mammalian brains and are used to determine common patterns that can be used to categorize data into certain classes and can be performed in very little time for large datasets. The ANN is trained using data known to be classified as “normal” and “abnormal”. Once trained, the ability of the algorithm to properly classify data is tested using the remaining spectra.

The specific parameters for the ANN for the oral cavity and the esophagus were different, and will be discussed specifically in Chapters 3 and 4 respectfully, but in general a select number of wavenumber inputs, dependent on differences in the two spectral classes, were passed through the nodes within the hidden layer and a binary output was produced as either normal or abnormal. Figure 2.15 shows an example of the ANN design for the oral pre-clinical trial data.
Figure 2.15 ANN design for the oral pre-clinical trial data classifying spectra as either “normal” or “abnormal”. Figure reprinted with permission from Mazur [23]

2.6 Papanicolaou Staining

Staining protocols were undertaken after data collection and analysis was completed in order to correlate results from SCP with a diagnosis from a cytologist. The method of the Papanicolaou “Pap” stain was used and the following is the step by step procedure utilized by the Laboratory for Spectral Diagnosis to stain all cellular samples. [24]

Step 1 - 20 dips in Tap Water
Step 2 - 3 minutes in Hematoxylin 1 [Richard-Allan Scientific, Kalamazoo, MI, USA]
Step 3 - 10 dips in Tap Water
Step 4 – 15 dips in Clarifier 1 [Richard-Allan Scientific, Kalamazoo, MI, USA]
Step 5 – 10 dips in Tap Water
Step 6 – 10 dips in Bluing Reagent [Richard-Allan Scientific, Kalamazoo, MI, USA]
Step 7 – 10 dips in Tap Water

Step 8 – 20 dips in 95% Ethanol

Step 9 – 2 minutes in Orange G-6 [Fisher Scientific, Agawam, MA, USA]

Step 10 – 10 dips in 95% Ethanol

Step 11 – 10 dips in 95% Ethanol

Step 12 – 3 minutes in Eosin Azure-50 [Surgipath Medical Industries, Richmond, IL, USA]

Step 13 – 10 dips in 95% Ethanol

Step 14 – 1 minutes in 95% Ethanol

Step 15 – 20 dips in 95% Ethanol

Step 16 – 20 dips in 95% Ethanol

Step 17 – 20 dips in 100% Ethanol

Step 18 – 20 dips in 100% Ethanol

Step 19 – 20 dips in Xylene

Step 20 – 1 minute or until ready to cover slip in Xylene

After removal from Xylene in the final step of the Pap stain protocol, the slides were allowed to dry and then a glass cover slip was applied using Permount toluene solution [Fisher Scientific, Fair Lawn, NJ, USA]. Slides were set to dry in the desiccator until ready for visual images to be collected. Images were collected using a QImaging GO2 3MB digital color camera attached to an Olympus BX40 microscope. An example of an image of a cellular sample collected after Pap staining is presented in Figure 2.16.
Figure 2.16 (A) Visual image of a cellular sample stained following the Pap stain protocol [15] collected using (B) a QImaging GO2 3 MB digital color camera attached to an Olympus Bx40 microscope.

2.7 References


[12] Bassan, P.; Byrne, H.J.; Lee, J.; Bonnier, P; Clarke, C; Dumas, P; Gazi, E; Brown, M.D; Clarke, N.W; Gardner, P. Reflection contributions to the dispersion artifact in FTIR spectra of single biological cells. *Analyst*, 2009, 134, 1171-1175.


CHAPTER 3: ORAL PRE-CLINICAL TRIAL

3.1 Introduction

3.1.1 Oral Disease

The American Cancer Society has published estimates for the rates of oral cavity (the mouth) and oropharyngeal (the area behind the mouth in the throat) cancer that the United States will see in 2013. These estimates predict that 36,000 people will be diagnosed with one of these cancers, 6,850 of which will succumb to the disease. [1] Data and statistics provided by the American Cancer Society will be for both the oral cavity and the oropharynx, while the research in this thesis will focus on the oral cavity.

Although there are multiple types of cells that make up the oral cavity and oropharynx, the most common types (9 out of 10) of cancers that develop in the oral cavity are squamous cell carcinomas. Oral cavity and oropharyngeal cancers are two times more common in men than women and more common in individuals over 55. There has also been a decrease in the death rate of these cancers in the previous 30 years. Countries that have a higher incidence of these cancers than the US include Hungary and France, while Mexico has a lower rate than all three. While the causes of these cancers are not well understood, some of the risk factors influencing the conception of these cancers are tobacco and alcohol usage, and human papillomavirus (HPV) in younger patients. [1-3]

The diagnosis of oral cavity and oropharynx cancers is often difficult because these cancers do not develop symptoms until they have reached more advanced stages. Once the cancers have reached these advanced stages, they often metastasize to other areas within the body. Some of the symptoms related to the development of oral or oropharyngeal cancer are
unhealing sores, bleeding lumps, white or red patches, swelling, or numbness. Biopsies are classically taken to diagnose whether a patient does indeed have a form of oral cavity or oropharyngeal cancer. After being diagnosed with oral or oropharyngeal cancer, a patient may be treated by surgery, radiation therapy, chemotherapy, targeted therapy, or palliative treatment. [1]

3.1.2 Oral Pre-Clinical Trial Background

The oral pre-clinical trial that was conducted by the Laboratory for Spectral Diagnosis began in 2010 after previous research showed that SCP was a reliable method for oral disease diagnosis. [3] The goal was to generate larger databases of “normal” samples to monitor the influences of tobacco use, medication, HPV vaccines, viral infections and history of abnormal oral diagnoses.

The Northeastern University IRB approved the pre-clinical trial so volunteer samples could be collected from members of the Northeastern community. Samples were collected from 93 patients from the tongue, cheek and mouth floor. Volunteers were asked to fill out an informational questionnaire developed to retrieve information such as gender, race, tobacco and medication usage, and oral cytological history.

Volunteer samples for the oral pre-clinical trial were collected, processed and their IR data were collected by Dr. Jennifer Schubert, Dr. Antonella Mazur, Dr. Ellen Marcsisin, and Kostas Papamarkakis. Subsequent data processing using the SkinnyCells program, described in Chapter 2, and analysis of data by PCA were performed by Kathleen Lenau.
3.2 Results

3.2.1 Previously Published

The oral cavity has been investigated in the past with optical spectroscopy as a diagnostic tool for oral cancer. [2] One group used a miniaturized fiber optic confocal reflectance microscope to perform *in vivo* imaging of oral neoplasia [4], and another group developed a low-cost multimodal, portable screening system for the early detection of oral cancer. [5]

Preliminary research at the Laboratory for Spectral Diagnosis to develop an early diagnostic method for oral cancers was first published in 2010. [3] Following these results the pre-clinical trial was developed in order to generate a method for the pre-screening of oral cancers and to investigate various influential factors.

Recent results for the pre-clinical trial were published by Dr. Antonella Mazur in her PhD dissertation entitled *Spectral pathology: automated classification of cytological & histological specimens utilizing infrared micro-spectroscopy.* [6] The pre-clinical trial was able to increase the number of samples for the normal database, which was composed of volunteers who did not take any medication, did not use tobacco, had no recent or active cold sores, no history of mononucleosis, no history of *herpes simplex*, no abnormal oral history, and had not received the Gardisil vaccine. There were a total of 19, 20, and 17 volunteers who were defined under this “normal” database for the cheek, mouth floor, and tongue, respectively.

All PCA results, those published previously by Dr. Antonella Mazur, and those presented for the first time in this thesis, were generated by randomly selecting the same number of spectra for each class. This was done in order to have comparative numbers of spectra in each class due to the large number of spectra in the normal database when compared to the much smaller
Dr. Mazur’s results mimicked those published in 2010 for differentiating cells from different regions of the oral cavity. SCP was able to distinguish between the three regions investigated for the oral pre-clinical trial. The differentiation between these three regions has been attributed to more collagen expression in the cells of the mouth floor and more keratin expression in the cells of the tongue. [3] From these results all following results were separated by region to remove the region as a possible variable.

As the end goal of this research is to create a diagnostic tool for oral disease the first results generated were to compare the normal samples and those clinical samples supplied by Tufts Medical Center that had been collected from areas of abnormal cytology, but appeared morphologically normal. The PCA scores plots showing the differences observed for the clinical and normal samples for the cheek, tongue, and mouth floor are presented in Figure 3.1.
The loading vectors, which represent the variance between classes, were plotted to observe the distinct spectral differences between the normal and clinical classes, and are shown in Figure 3.2. When this was done there was an obvious similarity noted between the first loading vectors for the three regions, which represent the largest variance between classes. It has been postulated that this commonality between the first loading vectors may be attributed to the disease manifesting within the cells. This variance appears to manifest within the protein regions.
of the spectra where the largest peaks are observed in the amide I and amide II locations, between 1800-1500 cm\(^{-1}\). The second and third loading vectors also saw this similarity, but not to such a large extent.

**Figure 3.2** The first principal component from each of the PCA scores plots that were presented in Figure 3.1, stacked to show their similarity, with the tongue (blue), mouth floor (red) and cheek (green). Figure modified with permission from Mazur [6].

The PCA plots in Figure 3.1 show that SCP can differentiate between normal and clinical samples. Next, an ANN was developed as a supervised trained algorithm to diagnose spectra as either normal or abnormal for the tongue, as it had the highest number of clinical spectra. The ANN, shown in Figure 3.3, was trained using 600 spectra, 300 randomly selected from the normal database, and 300 randomly selected from the clinical database. This ANN selected the 50 most different wavenumber points in the spectra containing 451 data points. A hidden layer was set to contain two nodes and an output neuron diagnosed the spectra as either normal or
abnormal. Following training of the ANN it was tested with a data set of 400 normal and 400 clinical spectra. The results from the ANN indicate that it has a very good specificity (94.3%), sensitivity (96%), and overall accuracy (95.1%).

![Confusion Matrix]

**Figure 3.3** ANN results diagnosing tongue spectra as either normal or abnormal with corresponding sensitivity and specificity. Modified with permission from Mazur [6].

Finally contributions from oral contraceptives in the cells collected from the mouth floor, and tobacco for those cells collected from the cheek were investigated. This was done to begin observing how everyday lifestyle choices could influence the cellular spectra of the different regions of the oral cavity. These results will be presented in Section 3.2.2 with additional results that have recently been generated for the remaining two regions.
3.2.2 Recent Results

The recently published results [6] discussed in the previous section also included results for tobacco in the cheek cells, and oral contraceptives in the mouth floor cells. As a continuation these same factors were investigated, but for the regions of the oral cavity not previously explored, and the results compiled.

From the volunteers sampled for the pre-clinical trial there were 10 who indicated that they were smokers, and had no other influences, such as medication usage, or history of mononucleosis, abnormal oral cytology, or herpes simplex virus, etc. The previous results comparing the smoker class to normal cheek samples showed interesting separation of a few subjects from the rest of the smoker class. New results showed very slight separation between the normal and smoker classes in the tongue and mouth floor regions of the oral cavity. The previous and newest results are presented together in Figure 3.4. These small variations may indicate that smoking does not cause a distinct change in the chemical makeup of a cell that is able to be detected within the spectra by SCP. However, some variables for the smoking class include the time since last cigarette, brand of cigarette, and lifetime smoking history. These variables may be confounding the results, and as such more information would be needed from the volunteers to properly analyze results.
Figure 3.4 PCA scores plot comparing the normal (blue) and smoker (red) classes of volunteers for (A) the cheek (B) the tongue and (C) the mouth floor. Results from the cheek were modified with permission from Mazur [6].

There were 3 volunteers who indicated that they had taken oral contraceptives within 24 hours of sample collection, and had no other influences, such as tobacco usage or history of mononucleosis, cold sores, or herpes simplex virus, etc. Previous results comparing the class of volunteers who indicated using oral contraceptives versus the normal class were shown for the mouth floor. [6] The PCA scores plot for the mouth floor, as well as those for the cheek and tongue are presented in Figure 3.5. In the PCA scores plots there are very distinct differences.
between the normal and oral contraceptive classes, which are likely attributed to the presence of known estrogen receptors in the mouth allowing estrogen from oral contraceptives to be expressed spectrally.

Figure 3.5 PCA scores plot comparing the normal (blue) and oral contraceptive (red) classes of volunteers for (A) the cheek (B) the tongue and (C) the mouth floor. Results from the mouth floor were modified with permission from Mazur [6].

Although it was inquired as to what type of medications the volunteers had taken within 24 hours of sample collection, it is unknown when exactly they had ingested the medication. Figure 3.6 shows the results comparing the normal class of volunteers to those five volunteers.
who had indicated that they had only consumed ibuprofen, Motrin, Tylenol, Advil, or aspirin within 24 hours of sample collection. These drugs are able to be taken approximately every 4 hours, which would imply that they are metabolized by the body once that time has elapsed and their spectral signature is no longer observed within the cells. Also, the dosage of medication or specific type may affect the results as well, resulting in the only slight separation between classes in the PCA results.

Figure 3.6 PCA scores plot comparing the normal (blue) and aspirin (red) classes of volunteers for (A) the cheek (B) the tongue and (C) the mouth floor.

SCP was shown previously to be able to detect the signature for the viral infection herpes simplex within the oral cavity. [3] For the pre-clinical trial there were only two volunteers who
had indicated that they had contracted the viral infection mononucleosis within the past 4/5 years. Figure 3.7 shows the PCA scores plots comparing the normal class with the two volunteers with a history of mononucleosis. Although the number of samples for the mononucleosis class is small, there are interesting separations observed in the PCA scores plots. The mononucleosis class in the cheek separated into two distinct clusters, one for each volunteer, with one cluster staying with the normal subjects, and one distinctly separating out. In the tongue and mouth floor the normal and mononucleosis classes show good separation between the two classes. However, the two volunteers who stated they had contracted mononucleosis within the past 4/5 years, also had a history of smoking, and had been taking other medications. It is unknown whether the observed separation could be contributed to the viral signature in the spectra, or to the other influences in the samples.
Figure 3.7 PCA scores plot comparing the normal (blue) and mononucleosis (red) classes of volunteers for (A) the cheek (B) the tongue and (C) the mouth floor.

3.3 Future Directions

The results presented in this thesis showed just a few of the influential factors that may affect spectra collected from cells within the oral cavity. Future goals are to explore the remaining influential factors for which volunteers indicated they were affected with, such as the Gardisil vaccine to protect against HPV, and the presence of cold sores indicating infection with the herpes simplex virus. To further investigate these factors and obtain accurate results it would
be necessary to increase sample volumes. Also, due to the number of variables that were found upon analyzing the results it would be beneficial to re-examine the volunteers and obtain more information, such as when exactly they had smoked, for how long they had been smoking, etc.

The final goal for this project would be to begin implementation in a hospital setting to use SCP as a diagnostic tool for oral disease. Before this can occur the method needs to be investigated with the different influential factors, and with more clinical samples with different diagnoses. With implementation in the hospital SCP could be used to monitor the progression of disease and assist cytologists with making accurate diagnoses.

3.4 Conclusions

The oral pre-clinical trial was able to complete the goals that were set forth upon its inception. A large database, consisting of 97 samples per region, was created for the normal oral samples, as well as the abnormal, everyday contributions from lifestyle choices. Preliminary results from the oral pre-clinical trial showed that SCP is a sensitive method able to detect the biochemical difference between cells from different regions of the oral cavity, those from smokers, with oral contraceptive use, and with oral disease. Further results showed that SCP could detect contributions from medications such as aspirin, and viral infections such as mononucleosis. Increasing the sample numbers has shown that SCP may be used as an early diagnostic tool for diagnosis of cancers of the oral cavity before the cells become morphologically different. The use of an ANN, showed that supervised training algorithms could be used as potential diagnostic tools for oral cancer.
3.5 References

CHAPTER 4: BARRETT’S ESOPHAGUS

4.1 Introduction to Barrett’s Esophagus

4.1.1 The Disease

Barrett’s esophagus is a metaplastic disease in which the normal squamous cells of the esophagus change to resemble the columnar cells of the intestine, as is illustrated in Figure 4.1. Barrett’s esophagus has been noted as a common risk factor for the development of esophageal adenocarcinoma (EAC), as patients with Barrett’s esophagus have a 100-fold increased risk of developing EAC. [1]

Following the development of Barrett’s esophagus, the disease may progress to low-grade dysplasia (LGD), then high-grade dysplasia (HGD), and eventually to EAC. [2] Most patients with Barrett’s esophagus do not progress past non-dysplastic disease or transient LGD. [3] Of the patients diagnosed with Barrett’s esophagus with HGD, 7.4% per year and 50% in 5 years will develop invasive cancer. [4] Once patients have progressed to full EAC, most survive less than a year after diagnosis. [5] In the United States, the incidence of EAC has dramatically increased, more so than any other malignancy in the past 25 years, with reports also showing a rise in the number of cases of Barrett’s esophagus worldwide. [6-8] Due to the asymptomatic nature of Barrett’s esophagus, a definitive diagnosis is often difficult to provide. [6]
Figure 4.1 (A) The cellular transformation that occurs in patients with Barrett’s esophagus represented in a diagram, reprinted with permission from Phillips, et al [1]. Visual images of (C) a normal squamous cell and (D) a columnar cell from a patient with Barrett’s esophagus.

The etiology of Barrett’s esophagus is not well understood, but typically the disease is associated with a history of reflux, in which gastric acid refluxes into the lower esophagus. [1] Patients who have a history of reflux are 10-fold or greater more likely to develop Barrett’s esophagus. [9] Although it is difficult to diagnose, reports indicate that Barrett’s esophagus is more likely to be seen in white males in the age range of 50-59 years. [10,11] Other risk factors for Barrett’s esophagus, although debated, include obesity, distribution of body fat, and smoking. [9,12,13] There has also been evidence of a genetic predisposition to the development of Barrett’s esophagus. [14] Various risk factors for the development of Barrett’s esophagus are
presented in Table 4.1.

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Direction</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male:female = 2:1</td>
<td>!!!</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Caucasian</td>
<td>!!!</td>
</tr>
<tr>
<td>Gastroesophageal Reflux</td>
<td>!!!</td>
<td>!!!</td>
</tr>
<tr>
<td>Adiposity</td>
<td>!!</td>
<td>!!!</td>
</tr>
<tr>
<td>Tobacco Smoking</td>
<td>!</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Dietary Fat</td>
<td>!</td>
<td>!!</td>
</tr>
</tbody>
</table>

**Table 4.1** Risk factors associated with the development of Barrett’s esophagus with corresponding agreement from literature as to that factor’s contribution to the development of disease and direction, which is the association of each factor. The greater the number of exclamation points, the higher the agreement and direction for that factor’s contribution to the development of Barrett’s esophagus. Modified with permission from Phillips, *et al* [1].

The mechanism for the cellular development of Barrett’s esophagus is also not well understood. Previous theories hypothesized columnar cells moving from the transformation zone into the esophagus; however this has been disproved. [15] Today, the more accepted theories are that the columnar cells come from the esophagus, rather than the transformation zone. There are multiple ways in which the columnar cells could develop within the esophagus, including an adaptive response during the healing of tissue from reflux. [1]

### 4.1.2 Current Diagnostic Methods

It is extremely important to develop early detection methods for Barrett’s esophagus as it is the leading cause of EAC. [16] Traditional methods for diagnosing Barrett’s esophagus are
often not able to detect the presence of Barrett’s esophagus until more advanced stages have been reached, so newer techniques have begun being investigated due to the increasing incidence of this disease. As with other cancers, patient outcomes can be significantly improved if the disease is detected and treated at an early stage. [17]

The progression from Barrett’s esophagus to EAC is classically monitored through endoscopic surveillance, which is undertaken approximately every 2 years. [12,16] Diagnosis is characteristically based on the grade or extent of dysplasia that is present. [12] During surveillance, biopsies from abnormal-appearing esophageal cavity may be retrieved and Barrett’s esophagus can be confirmed. [16]

Over the past decade, scientific advances have tried to catch-up with the rising incidence of Barrett’s esophagus in order to detect the disease before it progresses to EAC. Scientists have begun investigating the genetic markers, which may prove that Barrett’s esophagus is a predisposed condition. [18] There has also been an increase in the search for biomarkers that may assist in the diagnosis of Barrett’s esophagus and its progression to EAC. [18,19]

One report predicts optical techniques as an anticipated development for the recognition of dysplasia, and includes fluorescence, light scattering reflectance, Raman micro-spectroscopy, endomicroscopy, and optical coherence tomography. [16] Various spectroscopic techniques, such as SCP presented in this thesis, are beginning to be explored as diagnostic tools for Barrett’s esophagus. Raman spectroscopy has been shown to classify Barrett’s esophagus through the biochemical changes that occur in tissue. [20] One group performed in vivo Raman spectroscopy during a routine endoscopy, which took approximately 5 seconds to perform. Using semi-quantitative spectral modeling, they were able to observe changes associated to pre-neoplastic and neoplastic transformations in the gastric tissue. [21] Raman spectral mapping using PCA has
also been shown to probe the biochemical changes that are associated with malignancy in order to demonstrate these changes in the carcinogenesis of Barrett’s esophagus. [22] Endoscopic polarized scanning spectroscopy (EPSS) has also been used as a diagnostic tool for Barrett’s esophagus by providing a rapid scan system for the entire mucosal surface. [23] With this near real time method EPSS is also able to guide a biopsy by detecting and mapping sites to determine where precancerous dysplasia is located. [24]

4.1.3 Current Treatment Methods

The most common way to manage Barrett’s esophagus is through the treatment of reflux, which can also reduce the risk of transformation to EAC. [16] Treatment of reflux includes proton pump inhibitor (PPI) medication for the suppression of acid or surgery if symptoms continue after a medication trial. [16,25] Endoscopic surveillance, as mentioned previously, is also used to monitor Barrett’s esophagus to reduce the risk of cancer development, although there is debate about its usefulness and cost-effectiveness. [26-27] In the past, if the disease did progress to HGD, an esophagectomy was undertaken. [16]

A technique called ablation, referring to the destruction of the metaplastic cavity, is now more commonly used in place of a total esophagectomy, thus preserving the esophagus. Ablation can be conducted with multiple techniques including photodynamic therapy, argon plasma coagulation, cryotherapy, laser ablation, cryospray ablation, thermal ablation, and radiofrequency ablation. [13,16,25, 28-30]

Another technique known as endoscopic mucosal resection (EMR) is also used in place of an esophagectomy. In contrast to ablation, EMR is a technique in which the epithelium is excised, which is beneficial because it provides a definitive histological diagnosis. [28,31]
It has been shown that there is a possibility of reducing the risk of Barrett’s esophagus becoming EAC by treatment with aspirin and other non-steroidal anti-inflammatory drugs (NSAIDS) [19,25]. In addition, non-invasive drugs are being developed as a means of managing Barrett’s esophagus. [18]

4.2 Preliminary Results

Results were generated using 23 samples collected from 11 different patients received from the Department of Oesophagogastric Surgery in the Gloucestershire Hospitals NHS Trust. Of these samples, 10 had been diagnosed as being normal (2319 spectra), while 10 were diagnosed as having Barrett’s esophagus (1719 spectra), with 2 of the 10 collected from patients with visible Barrett’s who had dysplasia in the past (755 spectra).

A typical normal squamous cell measures approximately 40 μm in size, while a columnar cell measures 15 μm. Thus, the PapMap parameters were adjusted for a pixel size between 6-20 for the Barrett’s samples to include only the smaller cells with the hope of removing any normal squamous cells from the sample. The PCA scores plot which was generated utilizing the spectra for each diagnosis is shown in Figure 4.2. In this figure the spectra for normal squamous cells, cells from an area of Barrett’s esophagus, and possible dysplastic cells from an area of Barrett’s esophagus are represented by the blue, green, and red symbols, respectively.
Figure 4.2 PCA scores plot comparing (A) normal squamous cells (red), cells from an area of Barrett’s Esophagus (green) and possible dysplastic cells from an area of Barrett’s Esophagus (blue). (B) The same PCA scores plot shown in A, but with (i) the possible dysplastic class (ii) the Barrett's class, and (iii) the normal class made invisible.

The PCA scores plot seen in Figure 4.2 shows the spectral variance between the three classes of cells examined. The normal squamous cells appear to be the most different in the PCA scores plot and cluster separately from the other two classes. In order to view the differences more clearly, Figure 4.2 (B) shows the same PCA scores plot from Figure 4.2 (A), but with one class made invisible. In Figure 4.2(B)(iii), there does not appear to be an obvious separation between the cells from Barrett’s esophagus and those that may possibly be dysplastic. This may
indicate that the cells no longer contain the spectral signature for dysplasia, since these samples were taken from patients with visible Barrett’s who had a diagnosis of dysplasia in the past. Also, the differences between the normal and the two other classes may be so large that the small differences between the other two classes are not able to be observed.

The mean spectra for each of the three classes were compared to one another, shown in Figure 4.3, and the spectral differences were observed. The spectrum representing the normal class differs from the other two classes in the amide I and amide II regions. There are specific variations in the intensity of the bands between the normal and Barrett’s classes at 1713 cm\(^{-1}\), 1674 cm\(^{-1}\), 1522 cm\(^{-1}\) which are represented by the black stars in Figure 4.3. Also at certain wavenumbers there appears to be differences in the peak positions for the Barrett’s spectra and the normal squamous spectra, which are represented by the yellow stars. The differences between the three types of cells appear to be highly present in the protein region of the spectra, which may indicate that the disease is affecting the proteins within cells. There are also some slight differences in the lower wavenumber region around 1100-950 cm\(^{-1}\), which may be attributed to changes occurring in the DNA and RNA within the cells due to the onset of Barrett’s esophagus.
Figure 4.3 Mean spectra for normal squamous cells (red), cells from an area of Barrett's Esophagus (green) and possible dysplastic cells from an area of Barrett's Esophagus (blue). Black stars represent differences in intensity between classes, while yellow stars represent shifting in peak locations between classes.

The PCA plots in Figure 4.2 show that SCP can differentiate between normal squamous cells and cells that were collected from areas of Barrett’s esophagus. Following this observation, an ANN was developed as a supervised training algorithm to diagnose unknown data sets. Figure 4.4 shows the results obtained from the ANN.
Figure 4.4 (A) ANN results for the esophagus diagnosing biochemical differences as either normal or Barrett’s esophagus and (B) the outputs of this ANN.

The ANN was trained using 1200 spectra; 600 randomly chosen from the normal database and 600 randomly chosen from the Barrett’s database. The ANN was set to choose the 80 most different intensity points between the two classes of spectra. A hidden layer with three nodes and two output neurons was chosen. Once the ANN had been trained, it was then tested using a data set consisting of 1800 spectra. Again spectra were randomly chosen from what had not been used to train the ANN, but this time 900 spectra were randomly selected from each of the normal and Barrett’s data sets. The results from the ANN, shown in Figure 4.4, indicate a very good specificity (92.6%), a decent sensitivity (53%) and a good overall accuracy (77.8%) to successfully detect the differences between the normal and Barrett’s data. Sensitivity represents
the number of positive spectra (Barrett’s), which were correctly identified as being positive. The samples taken from patients with Barrett’s contained a mixture of Barrett’s columnar cells and normal squamous cells. The low sensitivity is most likely attributed to the presence of the normal squamous cells in these samples.

4.3 Future Directions

There are still tasks which must be completed for the Barrett’s esophagus project. In the future images of cells will be sent back to from the Department of Oesophagogastric Surgery in the Gloucestershire Hospitals NHS Trust where a cytologist will provide individual diagnoses. From here the spectra for cells that actually are classified as Barrett’s esophagus can be analyzed further to determine the differences between these cells and normal squamous cells spectrally. Better ANN results could be provided by training the ANN using spectra from cells known to have Barrett’s esophagus.

It is also a goal to increase the sample volume significantly. Also, as the one sample from the stomach and one with EAC did not provide enough cells for processing, increasing the sample volume for these areas would improve the reliability of results. It is especially important to increase the number of EAC samples so that the transformation from Barrett’s esophagus to cancer can be analyzed and better understood on a spectroscopic level.

4.4 Conclusions

Barrett’s esophagus is a pre-cancerous disease of intestinal metaplasia in which the lining of the esophagus changes to more closely resemble the lining of the intestine. The leading cause of Barrett’s esophagus is gastroesophageal reflux. Without proper treatment, Barrett’s esophagus
may progress to LGD and then to HGD, and then finally to EAC. Endoscopic surveillance must be undertaken every 2 years to visualize and biopsy the mucosa of the esophagus in order to detect early dysplastic lesions before progression to invasive cancer. SCP may represent an accurate and objective diagnostic tool which could potentially identify patients with Barrett’s esophagus by analysis of esophageal cells retrieved using minimally invasive sampling techniques, thus preventing the need for routine endoscopy and endoscopic biopsy. The results presented in this chapter show that SCP was able to differentiate between normal squamous cells and cells that were collected from an area diagnosed as having Barrett’s esophagus. The supervised technique of ANNs was also shown to have potential as a diagnostic tool for Barrett’s esophagus.

4.5 References


APPENDIX A: REPRINT PERMISSIONS FOR FIGURES

Permission to use figures in this thesis from previously published literature were obtained prior to use. Each of these permissions is shown here in the order they appear in the text.

Reprint Permission for Figure 1.1, 2.10, 2.13:

Email permission was received from Dr. Max Diem for the use of his figure in this thesis as Figures 1.1, 2.10, and 2.13.

“To Whom It May Concern,

I grant Kathleen Lenau permission to use/modify Figure 2 from my journal article entitled, "Applications of Infrared and Raman Microspectroscopy of Cells and Tissue in Medical Diagnostics: Present Status and Future Promises" published in Spectroscopy: An International Journal, in 2012, as well as my hand made figures.

Max Diem, Ph.D.”

While Figures 1.1 and 2.13 were provided directly from Dr. Max Diem, Figure 2.10 was previously published in: Diem, M.; Miljković, M.; Bird, B.; Chernenko, T.; Schubert, J.; Marcsisin, E.; Mazur, A.; Kingston, E.; Zuser, E.; Papamarkakis, K.; Laver. Applications of Infrared and Raman Microspectroscopy of Cells and Tissue in Medical Diagnostics: Present Status and Future Promises. N. Spectrosc. 2012, 27, 463-496. The publisher allows for the authors to retain copyright of their articles, as shown in the quote from Open Access, http://www.hindawi.com/oa/.

“Open Access is a publishing model that provides immediate, worldwide, barrier-free access to the full text of research articles without requiring a subscription to the journal in which these articles are published.

In this model, the publication costs are usually covered by the author’s institution or research funds. These Open Access charges replace subscription charges and allow the publishers to make the published material freely available to all interested online readers. At the same time, authors who publish in Open Access journals retain the copyright of their article.”
Reprint permission for Figures 2.15, 3.1, 3.2, 3.3, 3.4, 3.5:

Email permission was received from Antonella Mazur for the use of her figures. These figures were used/modified in the thesis for Figures 2.15, 3.1, 3.2, 3.3, 3.4, 3.5.

“To Whom It May Concern,

I grant Kathleen Lenau permission to use/modify Figures 3.26, 6.3, 6.4, 6.5, 6.6, and 6.8 from my Ph.D. dissertation entitled, "Spectral pathology: automated classification of cytological & histological specimens utilizing infrared micro-spectroscopy" published at Northeastern University in December of 2012.

Antonella I. Mazur, Ph.D.”

Reprint permission for Figure 4.1 and Table 4.1:

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

This is a License Agreement between Kathleen Lenau ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

<table>
<thead>
<tr>
<th>License Number</th>
<th>3094251118534</th>
</tr>
</thead>
<tbody>
<tr>
<td>License date</td>
<td>Feb 22, 2013</td>
</tr>
<tr>
<td>Licensed content publisher</td>
<td>John Wiley and Sons</td>
</tr>
<tr>
<td>Licensed content publication</td>
<td>Journal of Gastroenterology and Hepatology</td>
</tr>
<tr>
<td>Licensed content title</td>
<td>Barrett's esophagus</td>
</tr>
<tr>
<td>Licensed copyright line</td>
<td>© 2011 Journal of Gastroenterology and Hepatology Foundation and Blackwell Publishing Asia Pty Ltd</td>
</tr>
<tr>
<td>Licensed content author</td>
<td>Wayne A Phillips, Reginald V Lord, Derek J Nancarrow, David I Watson, David C Whiteman</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Licensed content date</td>
<td>Mar 21, 2011</td>
</tr>
<tr>
<td>Start page</td>
<td>639</td>
</tr>
<tr>
<td>End page</td>
<td>648</td>
</tr>
<tr>
<td>Type of use</td>
<td>Dissertation/Thesis</td>
</tr>
<tr>
<td>Requestor type</td>
<td>University/Academic</td>
</tr>
<tr>
<td>Format</td>
<td>Electronic</td>
</tr>
<tr>
<td>Portion</td>
<td>Figure/table</td>
</tr>
<tr>
<td>Number of figures/tables</td>
<td>3</td>
</tr>
<tr>
<td>Original Wiley figure/table</td>
<td>Figure 1, Figure 2, Table 1</td>
</tr>
<tr>
<td>Will you be translating?</td>
<td>No</td>
</tr>
<tr>
<td>Total</td>
<td>0.00 USD</td>
</tr>
</tbody>
</table>