OBSERVER AGREEMENT IN HISTOPATHOLOGICAL EVALUATION OF ORAL EPITHELIAL DYSPLASIA

by

Douglas James Brothwell

A thesis submitted in conformity with the requirements for the combined D.D.P.H./M.Sc. degrees, Graduate Department of Dentistry, in the University of Toronto

© Copyright by Douglas James Brothwell 1999
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-45970-5
Observer Agreement in Histopathological Evaluation of Oral Epithelial Dysplasia.


Douglas James Brothwell

Graduate Department of Dentistry, University of Toronto

Abstract

Objectives: We studied observer agreement in diagnosing oral epithelial dysplasia (OED).

Methods: Sixty-four slides were each examined twice by three oral pathologists. Information provided for each observation was altered in 50% of second examinations. The modal diagnosis was used as a gold standard to assess conformity.

Results: Pairwise interobserver weighted kappa (Kw) ranged from 0.68-0.83 with a group Kw of 0.74. Intraobserver Kw ranged from 0.82-0.96. Using a 3 point ordinal scale decreased Kw levels. Category specific simple kappa for normal, mild, moderate, severe and carcinoma-in-situ (CIS) were 0.51, 0.24, 0.26, 0.30, and 0.57. Information changing observer expectation significantly influenced diagnosis.

Conclusions: There was substantial interobserver consistency and almost perfect conformity when diagnosing OED. Agreement was best with CIS and lowest with mild dysplasia cases. Observer bias was present; consensus and calibration exercises may be appropriate. The full 5 point ordinal scale was best when assessing OED. Referring clinicians should include accurate patient information with specimens.
Acknowledgements

This study would not have been possible without the assistance of many individuals. I wish to acknowledge the efforts of the following three groups:

1. The pathologists and staff in the Department of Oral Pathology and Medicine, Faculty of Dentistry, University of Toronto. The considerable time and effort required to assess the histologic slide series was greatly appreciated. Thank you Drs. Bradley, Jordan, Leong, and Mock.

2. The members of my MSc. committee who volunteered their time and expertise to this study. The insights and advice were invaluable. Thank you Drs. Mock, Pharaoh, and Williams.

3. Finally, the co-supervisors of this study, my graduate education, and my community approach to health care. Your efforts and advice were greatly appreciated. Thank you Drs. Lewis and Leake.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abstract</strong></td>
<td>ii</td>
</tr>
<tr>
<td><strong>Acknowledgements</strong></td>
<td>iii</td>
</tr>
<tr>
<td><strong>Statement of Problem</strong></td>
<td>1</td>
</tr>
<tr>
<td><strong>Background</strong></td>
<td></td>
</tr>
<tr>
<td>Canadian Cancer Statistics</td>
<td>2</td>
</tr>
<tr>
<td><strong>Natural History of Oral Cancer</strong></td>
<td></td>
</tr>
<tr>
<td>Precancerous Lesions and Conditions</td>
<td>4</td>
</tr>
<tr>
<td>Oral Leukoplakia</td>
<td>4</td>
</tr>
<tr>
<td>Erythroplakia and Leukoerythroplakia</td>
<td>5</td>
</tr>
<tr>
<td>Oral Epithelial Dysplasia</td>
<td>5</td>
</tr>
<tr>
<td><strong>Secondary Prevention of Oral Cancer</strong></td>
<td>7</td>
</tr>
<tr>
<td>Assessing The Accuracy of Secondary Prevention</td>
<td>7</td>
</tr>
<tr>
<td>Comparisons Between Histopathologists</td>
<td>8</td>
</tr>
<tr>
<td><strong>The Assessment of Oral Epithelial Dysplasia</strong></td>
<td>9</td>
</tr>
<tr>
<td>Sources of Variation</td>
<td>9</td>
</tr>
<tr>
<td>Effect of Referral Information</td>
<td>10</td>
</tr>
<tr>
<td>a. Details of the clinical history</td>
<td>10</td>
</tr>
<tr>
<td>b. Expectation of prior prevalence</td>
<td>11</td>
</tr>
<tr>
<td>c. Referring Clinician's Tentative Diagnosis</td>
<td>11</td>
</tr>
<tr>
<td>Classification of Oral Epithelial Dysplasia</td>
<td>12</td>
</tr>
<tr>
<td>Treatment of Oral Epithelial Dysplasia</td>
<td>12</td>
</tr>
<tr>
<td>Need for Accuracy</td>
<td>13</td>
</tr>
<tr>
<td><strong>Agreement Terminology</strong></td>
<td>14</td>
</tr>
<tr>
<td>Validity/Accuracy</td>
<td>14</td>
</tr>
<tr>
<td>Bias</td>
<td>14</td>
</tr>
<tr>
<td>Reproducibility/Precision</td>
<td>14</td>
</tr>
<tr>
<td><strong>Assessing Agreement</strong></td>
<td>15</td>
</tr>
<tr>
<td>a) Validity &amp; Bias</td>
<td>15</td>
</tr>
<tr>
<td>b) Inter- and Intraobserver Agreement</td>
<td>15</td>
</tr>
<tr>
<td>c) Source of Disagreement</td>
<td>15</td>
</tr>
<tr>
<td>i. Within- vs Between-Observer</td>
<td>15</td>
</tr>
<tr>
<td>ii. Systematic vs Random Error</td>
<td>16</td>
</tr>
<tr>
<td><strong>A Gold Standard for Assessing Observer Agreement</strong></td>
<td>17</td>
</tr>
</tbody>
</table>
## Table of Contents

### Statistical Methods for Analysing Agreement
- Trend vs Concordance 18
- Agreement Indexes For Nominal and Ordinal Data 19
  - a. Percent Agreement 19
  - b. Dice's Concordance Index (Ps and Ps') 19
  - c. Goodman and Kruskal's Lambda 20
  - d. Rogret and Goldberg's $A$ 20
  - e. Chi-Square ($X^2$) 20
  - f. Phi 21
  - g. Pearson's Product Moment Correlation ($r$) 21
  - h. Spearman's Rho ($r_s$), Kendall's Tau (T) 21
  - i. Cohen's Kappa 21
    - i. Simple Kappa ($K_s$) 21
    - ii. Weighted Kappa ($K_w$) 22
    - iii. Multiple Observers and Kappa 23
    - iv. Paradoxical Behaviour of Kappa 23
    - v. Significance of Kappa 25
  - j. Coleman-Light's Conditional Agreement 26

### Sample Size in Observer Agreement Studies 27

#### Epithelial Dysplasia - Observer Agreement Studies 28
- a. General Magnitude of Agreement 28
  - i. Oral Epithelial Dysplasia 28
  - ii. Non-oral Epithelial Dysplasia 30
    - ii. a. Dysplastic Skin Nevi 30
    - ii. b. Dysplastic Colorectal Adenoma 31
    - ii. c. Cervical Dysplasia 31
  - b. The Effect of Lesion Severity on Agreement 33
    - i. Oral Epithelial Dysplasia 33
    - ii. Non-oral Epithelial Dysplasia 33
  - c. Inter- vs Intraobserver Agreement 34
    - i. Oral Epithelial Dysplasia 34
    - ii. Non-oral Epithelial Dysplasia 34
  - d. Effect of Pathologist Experience 34
    - i. Oral Epithelial Dysplasia 34
    - ii. Non-oral Epithelial Dysplasia 35
  - e. Number of Dysplastic Categories Utilized 35
    - i. Oral Epithelial Dysplasia 35
    - ii. Non-oral Epithelial Dysplasia 35
  - f. Effect of Referral Information 37
    - i. Oral Epithelial Dysplasia 38
  - g. Comparison Standards 39
    - Oral and Non-oral Epithelial Dysplasia 39
  - h. Effect of Observer Calibration/Consensus 39
    - Non-oral Epithelial Dysplasia 39

### Implications of Variability 40

### Study Purposes 41
Table of Contents

Material and Methods 42
  Sample Size 42
  Case Selection 42
  Inclusion/Exclusion Criteria 42
  Examiners/Examinations 43
  Study Design 43
  Statistical Methods 44
  The "Comparison Standard"
  Observer Blinding 46
  Ethical Consideration 46

Results 47
  1. General 47
  2. Extent of Observer Agreement 48
     2a. Presence/Absence of Dysplasia 48
     2b. 6 Point Ordinal Scale 48
     2c. 5 Point Ordinal Scale 49
     2d. 3 Point Ordinal Scale 49
  3. Category Specific Agreement 50
  4. Conformity with Modal Diagnosis 51
  5. Effect of Altered Prior Prevalence Expectation 51

Discussion 53

Tables
  Table 1 - Classification of lesions by signout and 3 observers 59
  Table 2 - Observer agreement - 2 point scale 60
  Table 3 - Observer agreement - 6 point scale 61
  Table 4 - Observer agreement - 5 point scale 62
  Table 5 - Observer agreement - 3 point scale 63
  Table 6 - Category specific group interobserver agreement 64
  Table 7 - Overall/category specific intraobserver agreement 65
  Table 8 - Conformity with modal diagnosis 66
  Table 9 - Effect of expectation of prior prevalence 67
  Table 10 - Comparison of 6 and 3 point scales 68
  Table 11 - Number of categories of discrepancy 69

Figures
  Figure 1 - Changes seen in epithelial dysplasia 70
  Figure 2 - Etiology of clinical disagreement 71
  Figure 3 - Example of perfect correlation, no concordance 72
  Figure 4 - Observer agreement 2x2 contingency table 73
  Figure 5 - Effect of prevalence on Kappa 74
  Figure 6 - Quantitative significance of Kappa 75
  Figure 7 - Sample size in observer agreement studies 76
  Figure 8 - Study design flowsheet 77

Appendices
  Appendix 1 - Summary of observer agreement studies 78

Bibliography 80
Statement of Problem

Tissue biopsy and microscopic examination of histological materials by oral pathologists is used to diagnose the presence and severity of dysplasia in oral mucosal lesions. The presence of oral epithelial dysplasia is a prognostic factor in the development of oral cancer. If dysplasia is identified, appropriate interventions may prevent the progression of these precancerous lesions. A necessary prerequisite to such preventive efforts is the accurate diagnosis of oral epithelial dysplasia. Diagnosis is more likely to be accurate when examiners agree with each other on the presence and severity of dysplasia.

Agreement between oral pathologists when diagnosing the presence and severity of oral epithelial dysplasia has been previously reported to be low. Unfortunately, the design and statistical analysis used by these studies has been inadequate to allow determination of the extent and source of the variation. Category specific agreement levels and the merits of using fewer diagnostic categories have not been previously reported for oral epithelial dysplasia.

Further, the interpretation of radiographs from various body locations has been shown to be influenced by providing the examiner with additional information at the time the radiograph is examined. The effect of providing oral pathologists with additional information at the time of histopathological examination of biopsied tissue has not been properly assessed.

The following study has been designed, in part, to address these deficiencies and answer the resultant questions.
Background

Cancer of the mouth is a serious condition. Just over half of afflicted individuals survive for five years (1). Oral cancer is malignant neoplastic disease, mainly squamous cell carcinoma (2), that occurs in the site grouping designated as 140-149 by the Ninth Revision, International Classification of Diseases. This group includes neoplasia of the lip, labial and buccal mucosa, gingiva, hard and soft palate, tongue, floor of the mouth, and pharyngeal areas. Although differing in etiology and prognosis, these sites are combined as a group due to their close anatomical association. Of these sites, the tongue is the most frequently affected; usually on its lateral border. The floor of the mouth, retromolar regions, and buccal mucosa are also known to be high risk sites (2). Recognizing the distribution of these high risk sites is important in the assessment of mucosal changes which may prove to be neoplastic or pre-malignant in nature.

Our understanding of oral cancer has improved in recent decades. Many cases of oral cancer are preceded by recognizable premalignant epithelial changes, the most important of which is thought to be the presence of oral epithelial dysplasia (3). The major risk factors for oral cancer are also known. They include tobacco use, alcohol abuse, and sun exposure (1,2). Diets adequate in iron, vitamin C, and vitamin A analogues may help protect against the development of oral cancer. With knowledge of the major risk factors and the premalignant dysplastic stage, primary and secondary prevention of oral cancer and its sequella is possible (3,4).

Canadian Cancer Statistics

Oral cancer is one of the cancer sites which has shown only a small reduction in incidence or survival in recent decades (5). For 1998, Statistics Canada estimated oral cancer would be the seventh most common type of cancer in Canadian males, sixteenth most common in Canadian females, and would afflict approximately 3,150 Canadians (5). An expected 1,040 Canadian oral cancer deaths result in an estimated deaths to cases ratio of 0.33 for 1998. This ratio is higher than that reported for breast, cervical, melanoma, or prostate cancers (5).
As with other types of cancer, oral cancer is not distributed evenly in the Canadian population. Over twice as many oral cancer cases and deaths occur in males as in females (5). This figure has changed from fifty years ago when males had five times the female oral cancer death rate (2). Like most carcinomas, the incidence of oral cancer increases with increasing age. In Ontario, the 1993 cancer registry showed that 91% of oral cancer cases occurred in the 32% of the population aged 45 years or older (6).

The age standardized incidence rates for oral cancer also vary by province in Canada (5). In males the 1998 estimated rate per 100,000 population ranges from a low of eleven in New Brunswick and Alberta to twenty-eight in Newfoundland. Corresponding 1998 estimated age standardized mortality rates per 100,000 males range from two in Saskatchewan to seven in Quebec and Prince Edward Island.
Natural History of Oral Cancer/Precancer

Squamous cell carcinoma comprises about 90% of oral cancer cases (1,2). Development of squamous cell carcinoma involves a step-wise series of changes in the genetic material which produce changes in the structure and behaviour of epithelial tissue (3). Eventually these changes become evident microscopically and clinically. Although some oral squamous cell carcinomas may develop directly from normal oral mucosa, others are preceded by a recognizable precancerous state. Thus oral cancer may develop as a two-stage process with the first step being development of a precancerous lesion. In the second stage, carcinoma develops within the precancerous lesion (7).

Precancerous Lesions and Conditions

The 1978 World Health Organization (WHO) working group defined a precancerous lesion as morphologically altered tissue in which oral cancer is more likely to occur than in its apparently normal counterpart (8). Similarly, they defined a precancerous condition as a generalized state associated with a significantly increased risk of cancer.

Oral Leukoplakia

Leukoplakia is the most common oral precancerous lesion (3). It is also the oral lesion most commonly showing dysplastic epithelial changes. Leukoplakia was defined by the WHO in 1978 as a white patch or plaque on the oral mucosa that can neither be rubbed off nor classified as any other disease (8). Its clinical appearance varies from a smooth, whitish area to a wrinkled, raised lesion. Leukoplakia is caused by the build-up of layers of keratin on the mucosal epithelium. It can present as keratosis only, but often shows degrees of disordered epithelial maturation and proliferation, both of which are features associated with premalignancy (3).

Oral leukoplakia is present in an estimated 3 to 4% of white Swedish and American adults (9,10). From 3 - 17% of these are diagnosed, at initial biopsy, as having carcinoma (11,12,17). Studies outside of India have reported malignant transformation rates for oral leukoplakia of between 4.4% and 17.5% when followed for up to 39 years (12-16). For leukoplakia which undergoes
malignant transformation, the transformation has been reported to occur an average of 2 ½ yrs after onset (17); therefore, early diagnosis is vital for these cases.

Research has shown that following tobacco use cessation, most oral leukoplakia will regress or disappear (16,18). Prospective studies have reported that if leukoplakia is diagnosed and appropriate primary prevention behaviours are adopted before cancerous transformation, the incidence rate of leukoplakia and the risk of oral cancer are decreased (19).

**Erythroplakia and Leukoerythroplakia**
The term erythroplakia is used analogously to leukoplakia to designate lesions of the oral mucosa that present clinically as red patches which cannot be characterized clinically or pathologically as any other condition (8). These lesions are usually irregular in outline, well defined, and have a bright velvety surface. The prevalence of erythroplakia is not known, but is recognized to be far less common than leukoplakia (3). Occasionally the red areas of erythroplakia may be intermingled with white patches of leukoplakia, producing what is called a leukoerythroplakia, or speckled, lesion. While leukoplakia is the most common premalignant lesion, erythroplakia and leukoerythroplakia carry a greater risk of malignant transformation (3,16,20,21) with up to 25.9% of cases progressing to cancer. Erosive leukoplakia is reported to show the highest transformation risk (16).

**Oral Epithelial Dysplasia**
Many factors are important in assessing the prognosis of precancerous lesions. These include location, non-homogeneous appearance, and presence of known risk factors. However, the most important prognostic indicator of malignant transformation to oral cancer is the presence or absence of epithelial dysplasia (22,23). Similarly, Kato et al. report that invasive cervical cancer is preceded by precursor lesions showing a spectrum of epithelial dysplastic changes (24).

Oral epithelial dysplasia (OED) is the diagnostic term used to describe the histopathologic changes seen in chronic, progressive, allegedly premalignant disorders of the oral mucosa. The
presence of epithelial dysplasia is an important indicator of malignant potential and the grade of dysplasia may be related to the likelihood of malignant transformation (25). Several precancerous lesions and conditions can show dysplastic changes (26) and may present clinically as leukoplakia, erythroplakia, or leukoerythroplakia (27). These conditions have been shown to respond to therapy (16,18,19).

Dysplastic epithelial changes are fairly common in the mouth. Besides leukoplakia, which has been reported to show dysplastic changes in from 16.7 to 24% of cases (11,16,25), dysplasia is found in other oral conditions including lichen planus, actinic cheilitis, nicotinic stomatitis, as well as several auto-immune conditions. Follow-up studies have shown that the presence of OED indicates an increased risk of developing invasive squamous cell carcinoma (16,25-31). One retrospective study of 308 OED diagnoses from 50,000 consecutive oral biopsies reported that the condition usually presents clinically as a white lesion (73.1%) in the floor of the mouth or the tongue (51.7%) (27). Excluding studies from India where different etiologic factors are thought to be important, research has shown an 11 to 36% malignant transformation rate of dysplastic lesions with study durations ranging from one to 40 years (19,25,27,29,31). In a large prospective study (670 patients), Banoczy reported a 13% transformation rate over 1 to 30 years (mean of 9.8 years) (16). Recently, Lummerman and colleagues reported that 16% of 44 dysplastic lesions became cancerous, with a mean transformation time of 33 months (27). Transformation rates as high as 36.4% have been described for certain types of dysplastic lesions (leukoerythroplakia) (15).

The relative proportions of different grades of epithelial dysplasia have also been reported. Lummerman et al. reported that of 308 consecutive acceptable biopsies signed-out as oral epithelial dysplasia, 59.4% were rated as mild, 27.6% moderate, 9.1% severe, and 4.0% carcinoma in-situ (27).
Secondary Prevention of Oral Cancer

Secondary prevention of oral cancer involves the early detection and treatment of lesions to reduce the mortality and morbidity associated with the disease. Early treatment helps prevent progression to advanced disease and death. With oral cancer, it has been shown that smaller (earlier) lesions have a better prognosis than larger lesions (1,32-34). Data show that patients with smaller and shallower lesions which have not spread to adjacent tissues or lymph nodes, have a better prognosis, need less invasive surgery, and cost less for the health care system to treat (3). Five year survival rates have been shown to be reduced by 33% in advanced disease as compared with earlier stages (33,34). Thus, when lesions are diagnosed early, prognosis is improved.

Many, if not all (30), cases of oral cancer are preceded by a recognizable premalignant dysplastic stage. It is for these patients that secondary prevention can have the greatest impact.

Assessing The Accuracy of Secondary Prevention

Early identification of lesions amenable to secondary prevention often involves screening programs where a test is applied to asymptomatic people in order to distinguish those that probably have the disease from those who do not. Screening examinations are not diagnostic; patients with a positive screening result are referred for diagnosis and necessary treatment.

Screening test accuracy is assessed by comparing the test result with a standard which, for the purpose of the study, has been defined as representing the truth. The chosen comparison standard, known as the 'gold standard,' varies according to the needs of the group being assessed. In a study of oral precancer screening in Sri Lanka, Warnakulasuriya and Pindborg compared the screening results reported by primary health care workers (midwives, public health inspectors, and nurses) with re-examination by a general dentist (35). In this study the diagnosis given by the general dentist represented the gold standard. Jullien et al. utilized the findings of a specialist dentist as the gold standard to assess the screening results of 24 general dentists (36). Similarly, Speight and colleagues utilized clinical examination by a specialist as the gold standard by which to assess the accuracy of diagnostic decisions made by general dentists (37).
Comparisons Between Histopathologists
Decisions regarding secondary prevention of oral cancer are largely determined by conformity with the comparison gold standard. Histopathological evaluation of biopsied tissue is considered to be the gold standard which most accurately represents the 'Truth.' There is, therefore, no gold standard available by which to assess the accuracy of histopathological evaluation. Its accuracy can only be judged by comparing the diagnoses of multiple examiners looking at the same case (interexaminer agreement), and by comparing the diagnoses made by the same examiner on repeat independent observations of the same case (intraobserver agreement).
The Assessment of Oral Epithelial Dysplasia

Although considered to be the most common premalignant lesion, most cases of leukoplakia do not exhibit epithelial dysplasia. When present, these dysplastic changes start in the basal epithelium, and extend to involve more of the epithelial thickness as the lesion becomes more dysplastic. The pathologist considers several morphological features when grading oral epithelial dysplasia. As summarized in Figure 1, cell differentiation, location of immature cells, cytonuclear features, hyperchromatism, nuclear to cytoplasmic ratio, pleomorphic nuclei, and location and appearance of mitotic activity are all considered to be important indicators of the dysplastic process (3).

Although attempts have been made to make the diagnosis of pre-malignancy more objective, studies of the grading of epithelial dysplasia in various human tissues have shown that agreement between different pathologists is less than perfect (22,24,25,27,38). As will be discussed later, the relatively poor observer agreement levels reported by some studies can be partially attributed to the use of bivariate data statistics for the analysis of observer agreement with ordinal epithelial dysplasia data. However, even with proper analytical methods, observer agreement levels remain less than perfect. Despite imperfect agreement, histological identification of OED in oral biopsies is the cornerstone of oral precancer surveillance.

Sources of Variation

Knowledge of the sources of observer variation may be helpful in improving observer agreement and ultimately patient treatment. The etiology of clinical disagreement as discussed by Sackett and colleagues is presented in Figure 2 (39). They suggest that variation can be categorized according to its source as being related to the examiner, the examined, or the examination. The radiology literature also reports many factors that affect the observation process (40-45). The following eight factors, incorporating factors common to both radiology and histopathology with those unique to histopathology, may be important in the histopathological diagnostic process:
1. The Examiner
   a. use of different cut-off thresholds
      i. observer training
      ii. observer bias
      iii. diagnostic criteria applied (different opinions on which histopathological characteristics are the most important for grading)
   b. expectation of prior prevalence
      i. information given to the observer at the time of the observation (ie., knowledge of the patient’s history, clinical description)
      ii. effect of a second opinion (ie., referring clinician’s provisional diagnosis)

2. The Examined
   a. biopsy taken from representative area of the lesion
   b. sections examined representative of the biopsy
   c. specimen properly handled and fixed
   d. histological preparation of good quality (thin section, properly stained)

3. The Examination
   a. clear optics and lighting
   b. specimens viewed in one or several sittings
   c. observer fatigue
   d. dividing a continuum of epithelial changes into an ordinal scale of classification.

Effect of Referral Information
Although all sources of disagreement were important, for the purpose of this study, only information given to the examiner was considered further. At least three types of referral information have been shown to influence observer performance in radiographic assessment (40-44). These factors may be important in evaluation of OED.

   a. Details of the clinical history
Studies have investigated the effect of presenting an appropriate case history to radiologists reading chest radiographs (40,43). Norman et al. reported that clinical histories affect ratings of
both diagnosis and of features present on difficult films (40). Similarly, Berbaum et al. reported improved detection of fractures when an appropriate history was provided (43).

For histopathological evaluation, microscopic evaluation ultimately establishes the presence and severity of epithelial dysplasia. However, as suggested by Krutchkoff et al., pathologists should be aware that clinical factors are helpful in overall diagnostic assessment (23). The authors suggest that tissues from some areas (ventrolateral tongue, floor of the mouth) and some individuals (smokers, drinkers, older individuals) should be viewed with greater suspicion than would similar findings in tissue from less susceptible sites and individuals.

The effect of providing clinical information to observers assessing OED has been studied (46). Contrary to expectation, agreement levels decreased when appropriate descriptive histories were provided to oral pathologists. The authors did not report whether the decreased agreement was statistically significant; nor were sufficient data provided to allow for the calculation.

b. Expectation of prior prevalence
The clinician's expectation of how likely a patient is to have disease has been shown to influence subsequent diagnostic measurements and decisions. For example, when studying the radiographic diagnosis of dental caries, Grondahl reported that observers were more likely to diagnose caries when told the prevalence was 75% than when told the prevalence was only 25% (42). This factor could be important in histopathological evaluation of OED, as the prevalence of dysplasia is known to be higher in smokers and the aged than in younger non-smoking individuals. Providing the patient's smoking status and age to the observing pathologist could potentially bias the diagnosis by altering the observer's expectation of prior disease prevalence (47). Expectation of prior prevalence may provide the mechanism by which knowledge of the patient history exerts an effect on subsequent diagnosis.

c. Referring clinician's tentative diagnosis
Berbaum et al. reported that when compared to not providing a tentative diagnosis, providing radiologists with an appropriate tentative referral diagnosis facilitated the detection of diverse chest lesions (43). The authors did not report the effect of providing an inaccurate diagnosis.
Classification of Oral Epithelial Dysplasia
The classification system used for oral epithelial dysplasia implies an orderly progression from normal tissue through to carcinoma; a view considered too simplistic by some authors (48). Nevertheless, the classification system does provide a useful foundation for management of these lesions. When assessing lesions, the pathologist initially categorizes lesions into a dichotomy based on the presence or absence of dysplasia. Dysplastic lesions are then further categorized into four ordinal categories based on the observed degree of dysplastic severity. Dysplasia is classified as mild if tissue changes are limited to the basal layers, moderate if changes extend as far as the middle of the spinous layer, severe if changes extend from the basal layer to above the midpoint of the epithelium, and carcinoma-in-situ (CIS) if the entire thickness of the epithelium is involved. Thus, the resultant five-point dysplasia grading scale consists of non-dysplasia and a four-point ordinal scale of increasing dysplastic severity (7).

Treatment of Oral Epithelial Dysplasia
The oral pathology report is a critical component in management of patients with intraoral lesions. While patient history and clinical suspicion may warrant the initial biopsy, it is the pathology report that largely directs future clinical actions (48). Although clinical treatment is influenced by the preference of the individual clinician, due to the malignant transformation potential of these lesions it is generally suggested that severe and carcinoma-in-situ cases be treated by complete surgical excision of the involved mucosa. It is also generally suggested that the changes found in mild and moderate dysplasia cases be reversed by the removal of causative factors. These cases are then followed prospectively, with further biopsies performed as early as three months later. Subsequently, the patient can be reviewed every 3 to 6 months and the lesion carefully assessed regarding the need for further biopsy (48).

Appropriate interventions for patients diagnosed with mild or moderate dysplasia may include stopping tobacco use, reducing alcohol consumption, maintaining good oral hygiene, consuming a diet adequate in fruits/vegetables, using anti-oxidant/retinoid/carotenoid drugs, topical chemotherapy treatment, and systemic antifungal drugs (50). Thus, the management of OED
usually consists of one of the three following options:

1) no treatment necessary (no dysplasia)
2) preventive therapy recommended with follow-up biopsies (mild or moderate dysplasia)
3) surgical therapy recommended (severe dysplasia or carcinoma-in-situ).

Most authors support this three point treatment schedule for both OED and cervical intraepithelial neoplasia (CIN) lesions (22,26,48,51-54). However, not all authors agree. For example, Bouquot, Abbey, and Lummerman each suggest that the diagnosis of any grade of OED usually mandates that the entire lesion be removed due to its malignant potential (9,27,55). Even if not surgically removed, the diagnosis of mild or moderate dysplasia usually results in treatment with non-surgical therapy and often requires additional future biopsies (48-50). Thus the pathology report is a key factor in future patient management.

**Need for Accuracy**

The mis-classification of dysplastic oral lesions can have serious health and future treatment implications for the patients involved. A false positive or higher-severity diagnosis could result in unnecessary surgical treatment for patients. Conversely, a false negative or lower-severity diagnosis could result in lesions with a high potential for malignant transformation remaining unexcised. Therefore, the accuracy and reproducibility of histological assessment of oral epithelial dysplasia is an important consideration in the management of patients with these lesions.
Agreement Terminology
Before discussing the assessment of observer agreement in histopathological evaluation of oral epithelial dysplasia, it is important to review the relevant terminology.

Validity
Validity is the degree to which observer measurement or diagnosis represents the actual situation of interest (56). The observer measurement or diagnosis is, therefore, compared with the "Truth." To assess validity in most situations, the observer measurement or diagnosis is compared with that obtained from a "gold standard," which, for the purpose of the study, is chosen to represent the truth. Observer accuracy in measurement or diagnosis, for most situations, is assessed by examining the degree of 'conformity' with a gold standard (57); this allows assessment of the 'correctness' of the measurement or diagnosis.

Bias
Bias is any systematic deviation of a measurement or diagnosis from the actual situation of interest (56). An example of observer bias is a situation where an observer rates every case of epithelial dysplasia one category higher than an accepted comparison standard.

Reproducibility
Reproducibility is the degree to which observer measurement or diagnosis remains the same on repeated independent observations of an unchanged characteristic (56). Rather than assessing correctness of a measurement or diagnosis, reproducibility considers the 'consistency' obtained on two or more observations of the same subject.
Assessing Agreement in Histopathological Materials

a) Validity and Bias
When assessing the histopathological evaluation of OED, there is no accepted gold standard appropriate to determine the validity of diagnosis. No other diagnostic test is considered to be more accurate than histopathological evaluation of biopsied tissue. Without a 'gold standard' for comparison, the true parameters for bias and validity cannot be determined. For this reason, when assessing the validity of OED diagnoses, conformity with a defined "comparison standard" is assessed (e.g. the modal diagnosis).

b) Inter- and Intraobserver Agreement
As stated previously, reproducibility is used to assess observer agreement. An interobserver reproducibility study involves a single assessment of all material by each of two or more raters blinded to each others' observation. The design is used to examine the interchangeability of examiner ratings. An intraobserver reproducibility design involves repeated independent assessment of a single subject by each rater. The design is used to evaluate examiner self-consistency. Combined, inter- and intraobserver agreement levels give an estimate of the degree of bias and validity present when an appropriate gold standard is not available.

c) Source of Disagreement
   i. Within- vs Between-Observer
Lack of agreement between observers can be due to within-observer variation or between-observer variation. Any apparent disagreement may stem from one or both of these sources. It is important to distinguish between these two sources of variation since the corrective action necessary to reduce or eliminate the variation depends on which type of variation dominates (58). Disagreement seen in interobserver study designs includes both within- and between-observer variation. Disagreement seen in intraobserver study designs is due only to within-observer variation. Combined, interexaminer and intraexaminer study designs allow an estimation of the ability to discriminate any given condition by any given examiner conducting a diagnostic test. They provide information regarding the degree of confidence that is appropriate for a given test
result, and may indicate the main source of variation as being within-observer or between-observers (58).

ii. Systematic vs Random Error
Disagreement among different examiners can be due to either systematic disagreement or random disagreement. Systematic disagreement occurs when different examiners consistently use different thresholds between diagnostic categories (bias). Examiners allocate different proportions of cases to each category, i.e., one examiner consistently assigns cases a higher score than another. This observer bias results in intraobserver agreement levels that are better than interobserver agreement levels. Random disagreement occurs when examiners are inconsistent in assigning diagnostic categories. The proportion of cases in each category does not necessarily change as one examiner grades the dysplasia higher in one case, but lower in the next (38). This inconsistency in assignment will be reflected by decreases in both intra- and interobserver agreement (57). From a practical viewpoint, systematic disagreement is more easily overcome. Since examiners are using different diagnostic thresholds, mutual education and case-sharing (calibration) can reduce the level of disagreement. Unfortunately, random disagreement can be very difficult to overcome.
A Gold Standard for Assessing Observer Agreement

Although no gold standard is currently available, it might be possible to assess the validity of histopathological evaluation of OED by defining a comparison standard. An appropriate comparison standard for assessing validity in histological assessment of OED could include at least one of the following: recorded sign-out diagnosis; expert panel consensus (55); the diagnostic mode (most common diagnosis among a number of independent examiners); or, long term follow-up of untreated cases to determine their natural history. Although the latter choice may provide the ideal comparison standard, due to ethical constraints, it is an unlikely choice. As well, the sign-out diagnosis is unlikely to prove an adequate gold standard, as it represents only a single diagnostic opinion. The remaining two options, consensus or modal diagnosis, could potentially be used as comparison standards.

It is important to stress that good intra- and interobserver reliability and good conformity with the modal diagnosis do not guarantee validity with a diagnosis. Rather, observed agreement levels provide an upper boundary on the degree of accuracy possible (59). Reliability is a necessary but not sufficient condition for validity.
**Statistical Methods for Analysing Agreement**

**Trend vs Concordance**

In describing the relationship between two variables, one can look at either trend or concordance (57). Trend is the strength of the tendency for changes in one variable to produce changes in the other variable. For example, as the water fluoride level increases in a community, the caries level will decrease in the population. Appropriate indices of trend would include the Pearson correlation coefficient \( r \), the regression coefficient \( b \), Spearman's Rho \( r_s \), and Kendall's tau \( T \) (57). These statistics can all be used to describe the relatedness, or mutual tendency, of two variables.

Concordance describes the extent to which two variables give the same result and, therefore, one could serve as a surrogate for the other (57). In research on the processes involved in diagnostic decision-making, we are usually more interested in concordance than in simple trend. Indices of trend are generally not adequate for investigating concordance since two variables can be very closely related, yet never agree. As shown in Figure 3, while two examiners may increase their severity rating in the presence of a certain clinical feature, one examiner could still rate consistently higher than the other. The two examiners could, therefore, show 100% agreement in trend, but no concordant diagnoses.

Potential concordance indices include: percentage agreement, sensitivity and specificity, Dice's coincidence index, simple and weighted Kappa, McNemar's test, and the intraclass correlation coefficient (57). Each of these indices has both advantages and disadvantages and the statistic of choice depends largely on the type of data available. The intra-class correlation coefficient (ICC) is the statistic of choice for the analysis of observer agreement with continuous data (60,61). However, since the consideration of oral epithelial dysplasia uses only nominal or ordinal data, the ICC will not be considered here.
Agreement Indices For Nominal and Ordinal Data

A variety of indices have been proposed for the analysis of observer agreement; the commonly used and most appropriate will be discussed here. The following formula and variables presented are based on the two by two contingency table format presented in Figure 4.

a. Percent Agreement

Percent agreement is the simplest and most frequently used index of agreement (57). It represents the level of agreement between observers as a percentage of all observations. Thus, for the two by two contingency table in Figure 4, percent agreement is calculated as:

\[ \% \text{ Agreement} = \frac{a + d}{T}. \]

Percent agreement is limited in its usefulness as a measure of observer agreement since it does not take into account the agreement between observers that would be expected to occur due to chance alone (59,60). Since chance agreement requires no diagnostic skill, indices that assess above-chance agreement are better indicators of diagnostic ability. For the two by two contingency table discussed above, total proportion of chance agreement is calculated by adding together the following two formula:

\[ \text{Chance agreement on presence} = \frac{R_1}{T} \times \frac{C_1}{T} \]
\[ \text{Chance agreement on absence} = \frac{R_2}{T} \times \frac{C_2}{T}. \]

b. Dice's Concordance Index (Ps and Ps')

Dice's concordance index is a measure of the probability that an observation by one observer will be repeated by another observer (62). Also known as the proportion of specific agreement, Ps and Ps' represent agreement with respect to presence and agreement with respect to absence, respectively (59). For agreement with respect to presence, Ps is calculated as

\[ Ps = \frac{2a}{R_1} + C_1. \]

For agreement with respect to absence, Ps' is calculated as

\[ Ps' = \frac{2d}{R_2} + C_2. \]

Ps and Ps' display the same shortcoming as percent agreement in that they do not correct for chance agreement between observers. When reporting kappa, Feinstein & Cicchetti suggest that Ps and Ps' should also be reported to improve understanding and to plan improvements in observer performance (63,64).
c. Goodman and Kruskal's Lambda ($\lambda$)

Goodman and Kruskal (1954) proposed an index of agreement that considers the frequency of correct predictions with and without knowledge of the joint ratings (59). Goodman and Kruskal's Lambda is calculated as:

$$\text{Lambda} = 2a - (b + c) / 2a + (b + c).$$

As with many measures of agreement, the principal problem with Lambda is that it does not correct for chance agreement between observers (59).

d. Rogret and Goldberg's $A$

To avoid the necessity of reporting the separate indexes $P_s$ and $P_s'$ for agreement on presence and agreement on absence respectively, Rogret and Goldberg proposed simply taking their mean (59). This value, $A$, is calculated as

$$A = \frac{1}{2} (P_s + P_s').$$

Again, $A$ falls short as an index of observer agreement in that it does not correct for chance agreement between observers. By combining $P_s$ and $P_s'$, $A$ also does not allow the independent consideration of the source of any observed disagreement (59).

e. Chi-Square ($X^2$)

Chi-square is a measure of association for categorical data that is used to test the null hypothesis that any relationship between observer agreement can be attributed to chance alone. Thus, Chi-square is a statistic which does correct for chance agreement (60). Chi-square is calculated as

$$X^2 = \sum [(O - E)^2 / E]$$

where $O$ = observed agreement
and $E$ = chance expected agreement

The use of $X^2$ as a statistic to assess observer agreement has several disadvantages. First, $X^2$ gives no indication of the strength or level of concordance. Second, as with most tests, it does not distinguish between a relationship of observer agreement and a relationship of observer disagreement; deviations from chance agreement in either direction contribute to the size of $X^2$.

Chi square is inadequate as an observer agreement statistic as it measures association of any kind, not specifically agreement. A chi square value which is significant at the 0.05 level indicates that
a statistically significant association exists between observer ratings, but says nothing about the
degree of agreement or the direction of agreement (57). Thus chi-square rates perfect
disagreement equal to perfect agreement.

f. Phi (Φ)
Phi is a measure of correlation between two dichotomous variables which is an index of the
strength of predictability of one variable from another. It is calculated as:
\[ \Phi^2 = \frac{X^2}{n} \]
As a measure of observer agreement, Phi suffers from the same shortcomings as \( X^2 \) in that it does
not distinguish a relationship of agreement from a relationship of disagreement (60).

g. Pearson's Product Moment Correlation (r)
Pearson's product moment correlation is an index of association used for interval data. It is an
index of how well observer B's observations are linearly related to observations by observer A.
The correlation \( r \) will equal 1 if all of observer A's and observer B's observations lie on a straight
line. However, the shortcoming of \( r \) as an index of observer agreement is that it does not account
for systematic observer bias (60). Thus, perfect correlation but zero agreement is possible (See
Figure 3). However, if combined with a Mann-Whitney test of mean rank differences, the
correlation coefficient can convey useful information.

h. Spearman's Rho (rs), Kendall's Tau (T)
Just as Phi is used to describe the trend between two dichotomous variables, non-parametric
correlation indices such as Spearman's Rho and Kendall's Tau have been applied to the analysis of
ordinal data. However, since they measure trend rather than concordance, they suffer from the
same limitations already described for Phi (57).

i. Cohen's Kappa

i. Simple Kappa (Ks)
Cohen's kappa, K, is the statistic of choice for evaluating observer agreement with dichotomous
data (60). Kappa is a chance corrected index which is calculated as
\[ K = \frac{Po - Pe}{1 - Pe}, \]
where \( Po = \) observed proportion of agreement = \( \frac{(a + d)}{T} \)
and \( Pe = \) expected proportion of agreement. = \( \frac{[(R1/T) \times (C1/T)] + [(R2/T) \times (C2/T)]}{T} \)
and therefore represents the observed agreement beyond chance as a proportion of the potential agreement beyond chance (57).

As discussed by Kramer and Feinstein, the null standard error (SEo) is used to test kappa for a significant difference from chance agreement (57), where

\[(SEo)^2 = Pe + (Pe)^2 - \sum [r_i c_i (r_i + c_i)]/n(1 - Pe)^2.\]

The standard error (SE) is used to calculate confidence intervals and to test the significance of differences between two independent values of kappa (57,60), where

\[(SE)^2 = A + B - C/n(1 - Pe)^2.\]

The explanation of this notation is found in Haas (60).

**ii. Weighted Kappa (Kw)**

Although appropriate for dichotomous data, simple kappa is not an appropriate statistic for assessing ordinal data. Simple kappa credits only exact observer agreement (in the main diagonal), and does not give credit for near misses in agreement (adjacent to the main diagonal). Thus simple kappa does not distinguish a one category disagreement from a four category disagreement. Although not as good as total agreement, a near miss is better than several categories of disagreement. Weighted kappa, Kw, is a modification of kappa intended to reflect the relative seriousness of disagreement between observers using ordinal classification. The advantage of Kw is that it uses a sequential weighting system that gives a greater penalty as the number of ordinal categories between two observers increases. Therefore, some credit is given to observers for near agreement (57). Due to the advantages offered by weighted kappa, it is considered to be the statistic of choice for evaluating observer agreement with ordinal data.

Weighted kappa is calculated in the same manner as simple kappa, except that each proportion in the 'observed' and 'expected' contingency table is multiplied by the appropriate weight from the chosen weighting system. Various weighting schemes have been developed for use with observer agreement studies, with quadratic and Cicchetti weights being commonly used (59). The Cicchetti weighting system was designed for use with combined dichotomous-ordinal scales like that used in assessing OED. By penalizing disagreements on the presence or absence of dysplasia
more severely than disagreements on severity grading, the Cicchetti weights are most appropriate for conditions where disease presence is an important treatment threshold. In the case of OED, the treatment commonly recommended for mild dysplasia consists of elimination of etiological factors such as tobacco use and alcohol abuse. Since these changes are also appropriate for patients without dysplastic lesions, the additional penalty imposed by the Cicchetti weighting system on discrepancies around the presence or absence of disease seems unjustified. For OED diagnosis, quadratic weights are most appropriate and will be used for reporting the results of this study.

iii. Multiple Observers and Kappa

Kappa is ordinarily used to measure concordance between two observers. When comparing three observers or more, there are two alternative methods of reporting kappa. Kappa indexes can be calculated for every possible pair-wise agreement between observers, i.e., A vs B, A vs C, and B vs C (57). These independent estimates of kappa can then be combined into an overall kappa ($K_o$) as described by Cohen (38), or by using the following formula (60):

$$K_o = \frac{\sum[K_i/V_i]}{\sum[1/V_i]}$$

where $V_i = (Se_i)^2$ is the variance of the $i^{th}$ estimate.

Alternatively, Fleiss (1981) has described a method for calculating a generalized agreement statistic for three or more observers that exploits common identities between kappa and the intraclass correlation coefficient through analysis of variance (59). Values for overall kappa are provided by the Walter’s software program PC-Agree (68).

iv. Paradoxical Behaviour of Kappa

Feinstein and Cicchetti have reported two paradoxes in the behaviour of kappa caused by imbalance in the contingency table’s marginal totals (63,64). In the first paradox, despite a high percentage of observer agreement, kappa can be drastically lowered by a substantial imbalance in the table's marginal totals either vertically or horizontally. This paradox is due to the correction built into kappa which gives a more severe penalty when the likelihood of chance agreement is higher and is usually caused when the majority of cases fall into a single row or column. Without a gold standard, the true prevalence of a condition cannot be determined. In this situation the
marginal distribution (row and column totals) act as surrogates for prevalence. If the surrogate 'prevalence' of disease is either very high or very low it creates imbalance in the contingency table's marginal totals (67). For example, if examiners agree that the prevalence is low prevalence is low, the majority of cases will fall into R2 and C2 of a two by two contingency table. The likelihood of chance agreement between two observers is therefore much higher, and is corrected for in determining overall observer agreement with kappa. The observer in these situations has fewer diagnostic decisions to make and observer agreement due to chance alone is more likely. As shown in Figure 5, when prevalence goes substantially above or below 0.5, the calculation of kappa becomes increasingly penalized for chance agreement.

The second paradox of kappa occurs when asymmetrical unbalanced marginal totals produce higher values of kappa than more symmetrical totals (63). This occurs when one examiner diagnoses most cases as positive (high prevalence) while a second examiner diagnoses the majority of cases as negative (low prevalence). The resultant asymmetrical distribution avoids the usual penalty imposed by kappa for high or low prevalence (63).

Cicchetti and Feinstein suggest that the best way to avoid paradoxical behaviour in kappa is to have balanced marginal totals in the 2x2 distribution (64). This can best be achieved, when studying oral epithelial dysplasia, by selecting even numbers of cases for each category. Even numbers can probably be obtained by using the original sign-out diagnosis in each case as a guide for case selection. However, this action will avoid the paradoxical behaviour of kappa only if the observers largely agree with the sign-out diagnosis. Only after all observations are completed will it be known if the 2x2 distribution is balanced symmetrically.

As reported by Cicchetti and Feinstein, calculating the values for Ppos and Pneg is important when kappa shows paradoxical behaviour (64). Ppos represents the observed proportion of positive agreement, and is calculated as

\[ P_{pos} = \frac{2a}{R_1 + C_1} \]

Pneg represents the observed proportion of negative agreement between observers and is
calculated as

\[ P_{\text{neg}} = \frac{2d}{R^2} + C_2. \]

\( P_{\text{pos}} \) and \( P_{\text{neg}} \) are identical to Dice's Coincidence Index described by Bulman (62).

As reported by Cicchetti and Feinstein, \( P_{\text{pos}} \) and \( P_{\text{neg}} \) make two important contributions when interpreting observer variability results (64). First, seeing \( P_{\text{pos}} \) and \( P_{\text{neg}} \) directly indicates the consistency on positive and negative diagnostic decisions (similar to sensitivity and specificity) and, therefore, indicate where and how improvements in observer agreement can be obtained. Secondly, \( P_{\text{pos}} \) and \( P_{\text{neg}} \) can help to explain the paradox of a high percent agreement but low kappa if it is present. If \( P_{\text{pos}} \) and \( P_{\text{neg}} \) show a large discrepancy in agreement for positive and negative diagnoses, the potential effect of a low kappa despite high percent agreement is immediately apparent.

v. Significance of Kappa

Kappa statistics can be assessed for two different types of significance. The first type is the quantitative significance shown by the kappa value obtained. Kappa values can potentially range from -1.0 to 1.0. A score of 1.0 indicates unity; perfect agreement has been achieved. A kappa score of zero indicates that observers have been classifying as if at random. A negative score indicates that the observers would have done better if they had just left things to chance, since they were diagnosing to different criteria (62). The level of quantitative significance for various kappa values is determined in a somewhat arbitrary fashion. The guidelines shown in Figure 6, as suggested by Landis and Koch, are commonly used (69). Although originally described for use with simple kappa values, these guidelines are suggested and commonly used for the interpretation of weighted kappa values (24, 51-53, 66, 76, 77).

The second type of significance for kappa is stochastic significance. It is a test of the null hypothesis indicating the proportion of results of this magnitude which could be expected to occur due to chance (i.e., \( P<.05 \)). It depends on the number of observations under study and on the variability which is present in the sample. A determination of whether kappa is significantly different from zero (the null hypothesis) is made by determining the z-value obtained by dividing
kappa by its null standard error $SE_0$ (57).

$$z = \frac{Kw}{SE_0}$$

It should be noted that there are two standard errors for kappa given in the literature, both of which were developed by Cohen and co-workers (59,60). The null standard error ($SE_0$) is used to test for a significant difference from chance agreement, while the standard error (SE) is used to calculate confidence intervals and to test the significance of differences between two independent values of Kappa. The z score used to calculate confidence intervals and to test the significance of differences between two independent values of Kappa is calculated as follows:

$$z = \frac{K1 - K2}{\sqrt{SE1^2 + SE2^2}}$$

Before these two standard errors were developed and published in 1969 (59), Cohen had suggested that there is little value in testing kappa for significance beyond chance, since concordance is normally expected to be better than chance (70). Since the hypothesis of agreement purely by chance is unrealistic in most circumstances, rejection of the null hypothesis does not provide useful information as the investigator expects agreement to be better than chance (58). When testing the significance of kappa, a large $n$ will cause small $K$ values to be significant while a small $n$ will cause large values of $K$ to be insignificant. However, a large value of kappa which is insignificant is indicative of a study design problem and not a concordance problem. For any observer agreement study with binary data, Cohen suggests that sample size of 30 will ensure significance for $K = 0.4$ (70).

**j. Coleman-Light's Conditional Agreement**

The Coleman-Light measure of conditional agreement is similar to $K$ in that it corrects for chance observer agreement, but differs in that the probabilities are calculated conditionally with respect to a particular rating of one of the observers (71). The Coleman-Light's conditional agreement is calculated for the 1st category as

$$K_1 = \frac{(T \times a) - (R1 \times C1)}{(T \times R1) - (R1 \times C1)}.$$ 

The weighted average of all the category specific $K_i$'s will equal kappa. The advantages and
disadvantages of Coleman-Light's conditional agreement index are identical to those already referred to for simple kappa (71).

**Sample Size in Observer Agreement Studies**

An important consideration in study design is ensuring an adequate sample size to produce the desired precision in estimation, while avoiding unnecessary examinations by cooperating clinicians. While sample size calculation methods for clinical trials and epidemiological investigations are well established, very little is written regarding sample size for observer agreement studies (58).

Freedman et al (1990), discussed the sample sizes necessary to study observer agreement with binary assessments (58). This is of limited assistance when studying observer agreement in the ordinal classification of epithelial dysplasia. Due to the lack of literature regarding appropriate sample size, most studies apparently based their sample size choices on practical considerations and in conformity with previous studies. Little consideration has been given to the required sample size and power necessary to achieve desired precision (22,24,38,49,72-81). The sample sizes for representative studies are presented in Figure 7.
Epithelial Dysplasia - Observer Agreement Studies

A summary of studies that evaluated observer agreement with histopathologic assessment of epithelial dysplasia is presented in Appendix 1.

a. General Magnitude of Agreement

i. Oral Epithelial Dysplasia

Three studies designed to assess observer agreement in diagnosing epithelial dysplasia were found; two additional studies were found which investigated other parameters and reported observer agreement data. Observer agreement in the assessment of oral epithelial dysplasia was first reported by Pindborg et al. at the 2nd meeting of the International Association of Oral Pathologists in Amsterdam (72). At that time, 72 oral pathologists independently submitted diagnostic suggestions based on nine photomicrographs displayed on a colour poster. A surprisingly wide spectrum of diagnostic suggestions resulted, ranging in two of the cases from no dysplasia to frank squamous cell carcinoma. Based on these findings, a call was made to develop an internationally accepted set of criteria for oral epithelial dysplasia. It is important to note that the conditions of the study did not reflect normal histological diagnosis. No slides were directly viewed and no other information (i.e., patient history) was given. Statistical analysis of the degree of variation was not reported, only raw numbers for each category were included in the article. Although the results are limited in their generalizability, the study importantly was the first reporting concern about observer agreement in histopathologic assessment of oral epithelial dysplasia.

Two articles published in 1995 investigated observer variation in histologic assessment of oral epithelial dysplasia. In the first, Karabulut and co-workers reported the results of four pathologists each examining 100 consecutive sections of oral leukoplakia (22). In all cases, information on sex, age, and lesion topography was provided. Interobserver percentage agreement ranged from 49% to 69% between the different pathologist pairs. Kappa values ranged from 0.27 to 0.45, showing poor to moderate agreement at best.

This study has several limitations. First, no assessment of intraexaminer variability was made so it
is impossible to estimate the degree of the observed variability that was due to within-examiner variation. Second, the 100 leukoplakia cases examined were consecutive cases and were not selected to give sufficient numbers of cases at each dysplasia grade. Consequently, the cases were heavily weighted at the low end of the severity scale (no dysplasia) and lightly weighted at the high end of the scale where improved agreement is generally shown. Thus the penalty imposed by kappa would be higher than if a more balanced distribution was used. Third, statistical analysis of the data was limited to reporting simple kappa values, which credit only perfect agreement.

The authors provided sufficient data for one pair of observers to allow the calculation of weighted kappa for this observer-pair. When a weighting system is applied to give credit for close agreement, the reported $K_s$ of 0.29 for observers A and B improves to $K_w$ of 0.47, more accurately reflecting the levels of observer agreement. This example points out the disadvantage of using a binary data agreement statistic when assessing agreement with an ordinal classification system.

In the second 1995 study, Abbey et al. had six oral pathologists each assess 120 oral biopsies exhibiting from simple hyperkeratosis to severe dysplasia (55). For interexaminer agreement, they reported simple kappa scores of 0.15 to 0.41 for exact classification agreement, and kappa scores of 0.29 to 0.57 for presence or absence of dysplasia. For intraexaminer agreement, they reported simple kappa scores of 0.05 to 0.49 for exact agreement, and 0.31 to 0.71 for presence or absence of dysplasia.

There are three points about the study design that should be noted. First, patient histories were not given, only histomorphology was used in judging the presence and severity of dysplasia. Second, rather than assess the consistency between different pairs of examiners, the study looked at conformity with an expert panel consensus comparison standard. Third, rather than the appropriate weighted kappa, only simple kappa statistics were reported. The authors attempt to compensate for excluding a weighted kappa analysis by calculating simple kappa values for agreement within one histological step. However, this is probably not appropriate as allowing for
one histological step on each side of the signout diagnosis creates a within-one-step category which is three categories wide. This simple kappa statistic for within-one-step, therefore, does not provide a penalty for being one category off.

In a retrospective study of malignant transformation of oral dysplastic lesions, Lummerman et al. reported observer agreement of only 54% between two pathologists (27). Similar to Abbey et al. the authors also report that their laboratory policy is to routinely recommend complete excision of all premalignant lesions.

When investigating use of low-dose isotretinoin (13-cis-retinoic acid) to prevent the malignant transformation of leukoplakia, Lippman et al. reported 92% agreement between two pathologists when assessing 38 randomly chosen specimens (28). Unfortunately, the study does not indicate if this agreement is for exact category agreement, or agreement on the presence or absence of dysplasia.

ii. Non-oral Epithelial Dysplasia
   ii a. Dysplastic Skin Nevi

Interobserver studies of dysplastic skin nevi are important to consider due to the similarities shown to oral epithelial dysplasia. There is evidence that dysplastic skin nevi are at least a marker and potentially a direct precursor lesion of melanoma. In one study, Duray et al. utilized a five point ordinal scale to assess observer variation (73). The scale combined the assessment of presence or absence of dysplasia, presence or absence of melanoma, and presence or absence of the two combined. It, therefore, does not attempt to categorize the degree of dysplastic severity. Kappa scores of 0.32 to 0.71 (mean = 0.46) were reported for the presence/absence of dysplasia. Although repeat examinations were performed for 50 slides, intraexaminer agreement was not reported. Kappa statistics were calculated for each of the five examiners, but only as conformity with a single reference pathologist.

Another study of melanotic nevi was more thorough in the statistical analysis of observer agreement. Piepkorn et al. studied observer agreement using a 3 point ordinal scale (74). They
reported an average intraobserver kappa of 0.63 and average interobserver kappa of 0.34. Kappa values were only calculated for the presence or absence of dysplasia. No attempt was made to assign a severity rating to the positive dysplastic cases.

Duncan et al. investigated the agreement among dermopathologists on the presence or absence of melanoma and dysplasia, and on categorizing dysplastic severity as slight, moderate or severe (75). The study reported kappa values ranging from 0.55 to 0.84 (mean = 0.68) for the five observers when assessing presence of dysplasia and melanoma. For the severity of observed dysplasia, kappa values dropped to from 0.05 to 0.47 (mean = 0.27). These low kappa values are partly explained by the use of simple kappa statistics only.

ii b. Dysplastic Colorectal Adenoma

Similar to oral epithelial dysplasia, dysplasia in colorectal adenoma indicates the risk of developing future colorectal cancer. In a study of 187 slides and three examiners, Jensen et al. reported simple group kappa values of 0.58 and 0.33 for intraobserver and interobserver agreement, respectively (66). When weighting was used to allow for near misses, the kappas improved to 0.67 and 0.42. The improved levels of agreement with the weighted kappa reflect the advantage of a less severe penalty when disagreement is small.

Using a five point ordinal scale of severity for anal intraepithelial dysplasia (AIN), Carter et al. reported simple kappas ranging from 0.09 to 0.48 (76). These scores improved to from 0.17 to 0.60 when weighted kappa was used.

ii c. Cervical Dysplasia

Ismail et al. showed generally poor to moderate levels of observer agreement especially in grading of low level cervical intraepithelial neoplasm (CIN) (77). They reported a group intraobserver simple kappa value of 0.55 and interobserver simple kappa value of 0.35. When using weighted kappa to allow for close agreement, these values increased to 0.85 and 0.77 respectively. Although the weighted kappa values represent generally good agreement, the authors were concerned about the low kappa values found for categories CIN 1 and CIN 2. They recommended changes in the terminology of CIN lesions to allow for only two categories: those
that require treatment and those that do not and can be followed up (77,78).

In a similar study, De Vet et al. reported an unweighted group kappa value of 0.28 and a weighted group kappa value of 0.56 (51). Reporting only the percentage agreement for variation in different categories, the authors suggest that all grades of dysplasia were equally difficult to distinguish from adjacent categories. Intraobserver agreement was not reported.

A follow-up study was done by De Vet et al. to assess the sources of the variation (52). Prior to the study, a consensus was reached on which morphological characteristics should be considered relevant for grading. They reported a weighted kappa score improved to 0.69. Paradoxically, the observers agreed better on the degree of dysplasia than on the morphologic characteristics on which the diagnosis was based. This finding is suggestive of 'pattern recognition' in the decision making process where pathologists first judge the degree of dysplasia, and then score the morphologic characteristics to match the diagnosis. This process of pattern recognition may be similar to that described for the diagnosis of dental caries (82).

When assessing the effect of adding an extra 'uncertain' category to the grading of CIN, Creagh et al. reported a simple group kappa value of 0.23 and a weighted group kappa value of 0.50 (53). Category specific simple kappa values ranged from 0.52 for CIN III to 0.06 for the new borderline category. The authors concluded that the addition of an extra category was not helpful and that the diagnosis of CIN was poorly reproducible. Intraexaminer agreement was not reported. Jones et al. chose not to report kappa statistics and relied exclusively on percent agreement (54). The study results, therefore, do not take chance agreement into account when reporting interobserver agreement levels of 87-100% and intraobserver agreement levels from 88.5 - 100%.

Kato et al. reported a weighted overall kappa of 0.56 for three histopathologists (24). They reported that simple kappa category specific agreement was best for CIN III (0.58) and invasive cancer (0.74) and worst for CIN I and II (0.12 and 0.16).
b. The Effect of Lesion Severity on Agreement

Although categorical differences in observer agreement can be assessed by use of category specific percentage agreement, it is preferable to use a chance corrected measure of agreement such as the Coleman Light index. With this index, observer agreement statistics are calculated conditionally with respect to a particular rating of one of the observers and, therefore, indicate agreement levels based on the rated severity of the lesion. Unfortunately, few studies achieve this level of sophistication.

i. Oral Epithelial Dysplasia

In the Karabulut et al. study, the 100 leukoplakia cases examined were consecutive cases and were not selected to give sufficient numbers of cases at each dysplasia grade (22). Consequently, the cases were heavily weighted at the low end of the severity scale (no dysplasia) and lightly weighted at the high end of the scale where improved agreement is generally shown. This may have added to the low observer agreement levels that were reported, although category-specific agreement levels were not reported.

Abbey et al. avoided this problem by selecting cases to give a proportion of dysplasia cases greater than typically seen in a biopsy service (46,55). The authors report that this was done to keep the sample size reasonable while ensuring that an adequate number of each type of epithelial dysplasia was used. Again, category-specific agreement levels were not reported.

ii. Non-oral Epithelial Dysplasia

Fenger et al. reported category specific kappa values in grading colorectal adenomas with two different classification systems (80). Kappa values reported for mild, moderate, severe, and CIS with the first system were 0.47, 0.51, 0.29, and 0.47 respectively. With the second system, kappa values reported for mild, moderate, and severe were 0.43, 0.21 and 0.80 respectively. These values indicate that the classification system used and the severity of the lesion may both be important factors in observer agreement.

When assessing cervical epithelial abnormalities, Ismail et al. reported almost perfect agreement
above chance for evaluation of invasive carcinoma, substantial agreement for severe dysplasia, moderate agreement for normal and reactive categories, and slight to fair agreement for mild and moderate dysplasia of the uterine cervix (77). Creagh et al. reported only slight or fair agreement in normal, mild, and moderate categories but moderate agreement in assessing severely dysplastic tissue (53). When assessing agreement between three pathologists, Kato and colleagues reported substantial agreement with invasive lesions (0.74), moderate for severe dysplasia (0.58), fair agreement with normal tissue (0.39), and only slight agreement for mild and moderate dysplasia (0.12 and 0.16) (24). Conversely, De Vet et al. reported that all grades of cervical dysplasia were equally hard to distinguish from adjacent categories (51).

c. Inter- vs Intraobserver Agreement

i. Oral Epithelial Dysplasia
Abbey et al. were the only authors reporting intraobserver agreement with oral epithelial dysplasia (55). When reassessing 60 selected slides, 3 pathologists showed better intraobserver agreement than interobserver agreement (agreed better with themselves than with other observers). The remaining 3 pathologists showed interobserver agreement levels that were better than their intraobserver agreements (agreed better with other observers than with themselves).

ii. Non-oral Epithelial Dysplasia
Generally, intraobserver agreement levels are better than those reported for interobserver agreement (66,74,77).

d. Effect of Pathologist Experience

i. Oral Epithelial Dysplasia
Karabulut and colleagues compared the histopathologic examinations of two oral pathologists with that of two general pathologists (22). They reported that the interobserver variability was due to individual differences rather than to educational background. When looking at simple kappa values, agreement between examiners with similar educational backgrounds was not significantly different than that between examiners with different educational backgrounds. Unfortunately, comparisons were not based on the appropriate weighted kappa statistic.
ii. Non-oral Epithelial Dysplasia

In a study of observer agreement in grading melanotic nevi, Duncan et al. reported that experienced dermatopathologists, with experience of between six and over twenty years, showed simple kappa values of from 0.38 to 0.47 (75). Inexperienced dermatopathologists, pathologists with less than 2 months experience in their hospital, showed simple kappa values ranging from 0.05 to 0.24. They concluded that the grading of melanotic nevi is a skill that can be learned through experience.

e. Number of Dysplastic Categories Utilized

i. Oral Epithelial Dysplasia

All identified studies of agreement in histopathological evaluation of oral epithelial dysplasia have used either 5 diagnostic categories (none/mild/moderate/severe/in-situ) (22) or 4 diagnostic categories (none/mild/moderate/severe) (27,28,55). The relative merits of different numbers of diagnostic categories has not been reported.

ii. Non-oral Epithelial Dysplasia

The number of diagnostic categories used varies in different studies. Duray et al. utilized only two categories, present/absent, when assessing agreement with dysplastic nevi (73), while Piepkorn and colleagues added an intermediate category when studying the same topic (74). Duncan et al. reported separate results obtained when assessing melanotic nevi on two different three-category scales, normal/dysplasia/melanoma, and mild/moderate/severe dysplasia. These results were previously reported.

Jensen et al. utilized only three categories, mild/moderate/severe dysplasia, when assessing colorectal dysplasia (66), while Carter et al. added normal and squamous carcinoma categories when assessing AIN on a five point ordinal scale (76).

Ismail et al. used seven categories when assessing cervical epithelial tissue (77). The large number of categories in this study demonstrate the importance of using weighted kappa to properly assess observer agreement. While a simple kappa value of 0.35 indicated only fair interobserver
agreement in this study, the weighted kappa value of 0.77 indicates substantial agreement. Similarly, while intraobserver agreement appeared to be only moderate ($K = 0.54$) when assessed with simple kappa, this agreement was shown to be very high ($Kw = 0.85$) when weighted kappa was used. Also assessing dysplasia of the uterine cervix, Creagh et al. assessed agreement when using a six ordinal category classification scheme (53) while De Vet et al. utilized a five point ordinal scale to describe his histopathological findings (51,52).

As mentioned previously, Carter et al. studied observer agreement in assessing anal intraepithelial neoplasia (76). Due to the unacceptably high level of observer disagreement, the authors recommend further study with a simplified classification system, dividing AIN into two grades, high and low. This approach has also been discussed by experts studying cervical epithelial abnormalities. Ismail et al. showed generally poor to moderate levels of observer agreement especially in grading of low level cervical intraepithelial neoplasm (CIN) (77). They recommended changes in the terminology of CIN lesions to allow for only two categories: those that require treatment and those that do not and can be followed up (77,78). Similarly, Epstein et al. reported the interobserver agreement in prostatic intraepithelial neoplasia (PIN) (79). They reported that the level of agreement was fair (kappa = 0.33) when 6 categories were used, while agreement was substantial (kappa = 0.61) when 3 groups were used.

While many investigators agree that use of fewer diagnostic categories is desirable due to the improved observer agreement, there is some dissent. Morris suggests that kappa statistics are ideal for use when assessing agreement with nominal data but inappropriate for ordinal data (49). He suggests that the use of kappa statistics in histopathology is problematic since it involves the ordinal classification of dysplastic changes which occur on a continuum. Concepts from information theory are recommended as a means of transmitting the maximum amount of information from histopathologist to clinician. Rather than reducing the number of categories to improve agreement, as measured by kappa statistics, information theory recommends increasing the number of categories to a 100 point scale which is quoted with 95% confidence limits.
One weakness is identified in Morris' argument. He states that disagreement between observers is caused by random categorization of cases which fall between the centre points of adjacent categories. If this were true, we would expect that any two observers would not disagree by more than one category on any given case. However, this has been shown to not be the case. With 9 cases of oral epithelial dysplasia, Pindborg et al. reported the closest agreement to be a case where pathologists disagreed by three categories in the diagnosis (72). With all other cases, the degree of disagreement was four, five or six categories. Similarly, Karabulut et al. reported the case categorization of two examiners (22). Of 67 cases rated as no dysplasia by examiner A, examiner B differed by 2 categories in 8 cases, 3 categories in 2 cases, and by 4 categories in 2 more cases. Obviously, the lack of agreement does not only derive from random categorization of cases which occur midway between adjacent categories. His discussion of the adequacy of the kappa statistic for assessing observer agreement is certainly true for the inappropriate use of simple kappa when assessing agreement with ordinal data.

f. Effect of Referral Information in Diagnosing Oral Epithelial Dysplasia
Histological assessment of oral epithelial dysplasia normally involves grading specimens according to the severity of dysplastic change. It is generally assumed that diagnosis, and the signs and symptoms used to derive the diagnosis, are distinct entities (40). Textbooks and journal articles describe the process of histopathological diagnosis as the independent consideration of patient history, referral description, and histopathologic features. Ignored is the possibility that provisional diagnostic decisions based on the patient's history and the referring clinician's description, can influence decisions made about which features are present (41-45). If this occurs, the grade assigned at diagnosis may be biased by knowledge of the patient's history and the referring clinician's description and clinical impression. This could represent a double weighting of the 'referral information' into treatment decisions made by the clinician. After receiving the pathological report which has been influenced by the 'referral information,' the treating-clinician considers the pathology report, patient history, and the referral lesion description when making treatment decisions. In this way, the referral information is double weighted in treatment decisions.
Studies of observer agreement in histopathological evaluation of epithelial dysplasia have addressed the issue of referral information in different ways. Some studies have provided the participating pathologists with the original signout referral description (27,28), while others gave an abbreviated history (22) and most gave no patient information at all (51-53,55,66,73,76,77).

In a recent study, Abbey et al. investigated the effect of supplying appropriate clinical information on the diagnosis of oral epithelial dysplasia. They reported a decrease in observer agreement when compared to the previous study where the same examiners had evaluated the same slides without clinical histories (46). The authors did not report whether or not this difference was statistically significant.

Although there has been little other research on the effect of 'referral information' in histopathological evaluation, it has been shown that the interpretation of radiographs is influenced by providing the reader with additional information at the time the radiograph is examined. In a study of the effect of accompanying clinical history on radiological interpretation, Norman et al. reported that clinical histories affected ratings of both the diagnosis and the features reported as present on diagnostically difficult radiographs (40). Thus, at least in ambiguous cases, diagnostic uncertainty may result in increased reliance on the referral information as an information source for decision making.

Berbaum et al. reported that providing the referring clinician's tentative diagnosis resulted in increased true positive diagnosis rates for detection of diverse chest lesions (43). The improved outcome presumably results from providing an accurate tentative referral diagnosis. Similarly, Schreiber reported improved true positive rates with a small increase in false positive rates when a clinical history was submitted with chest x-rays (44). It is anticipated that the patient's history and referring clinician's description will also influence decisions regarding histopathological evaluation of oral epithelial dysplasia.
g. Comparison Standards

**Oral and Non-oral Epithelial Dysplasia**

Typical comparison standards used to assess validity of histopathological assessment of epithelial dysplasia include signout diagnosis (73), a single expert pathologist (66), and panel consensus (22,55).

h. Effect of Observer Calibration/Consensus

**Non-oral Epithelial Dysplasia**

The effect of obtaining consensus on which characteristics are relevant for grading epithelial dysplasia was investigated by De Vet et al. (52). After an initial study showed poor agreement with no observer calibration, a second study was performed after reaching consensus on what features were to be considered for the grading of cervical dysplasia. They decided to score the following characteristics: location of immature cells; hyperchromasia; nucleus/cytoplasmic ratio; polymorphic nuclei; location of mitotic activity; and appearance of mitotic activity. As a result of the calibration exercise, group kappa values were reported to increase from 0.27 and 0.55 to 0.32 and 0.69 for simple and weighted kappa respectively. It is important to note that, although consensus was reached on which histopathological features were important to consider, no attempt was made to rank the importance of the various characteristics. Similar results would presumably be obtained following consensus and calibration exercises for the diagnosis of oral epithelial dysplasia.
Implications of Variability

Histopathological diagnosis and grading are the ‘Gold Standard’ in the diagnosis of many conditions. Clinical decision-making for the treatment of patients with precancerous oral epithelial lesions is largely directed by the pathology report. It is evident from the many studies of histopathological evaluation of epithelial dysplasia, that observers often show less than perfect agreement either with their own previous diagnoses or with those of colleagues. Since it is common to recommend surgical removal of lesions displaying severe dysplasia or carcinoma in situ, if not all dysplastic lesions (27,55), this mis-classification has significant ramifications on patient treatment. The imperfect agreement shown between histopathologists when evaluating these lesions necessarily produces under-treatment of some patients and over-treatment of others.

Previous studies of observer agreement in histopathological evaluation of oral epithelial dysplasia have identified that agreement is less than perfect. Unfortunately, due to weak study design and statistical analysis, they are of limited use in delineating the extent of observed disagreement or the source of the observed disagreement. The relative merit of using a reduced number of diagnostic categories to improve observer agreement levels has not been assessed with oral epithelial dysplasia. Finally, the 1998 study undertaken by Abbey et al. (46) raises, but does not answer, the question “What is the effect of expectation of prior prevalence on the diagnosis of oral epithelial dysplasia?”
Study Purposes

As discussed in the previous literature review, the subject of observer agreement in histopathological evaluation of oral epithelial dysplasia has been addressed in a few published studies (22,46,55). Whereas the histological materials and the participating pathologists were similar to this study, much remains to be determined about the extent and sources of variation in the diagnosis of OED. First, intraexaminer agreement was not reported in one study (22) and reported with only simple kappa statistics in others (46,55). Thus the extent of reported observer agreement has probably been underestimated and the relative importance of within- and between observer error remains unclear. Second, while studies of category specific agreement in diagnosing CIN have shown definite trends in category specific agreement, no results have been published for OED. Third, the relative merits of using fewer diagnostic categories in the categorization of OED has not been reported. This issue has received considerable attention for melanotic nevi, colonic, anal, and cervical dysplasia (75,76,77,80,86). Finally, while the expectation of prior prevalence has been shown to affect diagnostic decisions in radiographic interpretation (40,42-44) and may possibly exert an effect of the diagnosis of OED (46), its importance has not been fully elucidated in the histopathological evaluation of oral epithelial dysplasia.

In response to deficiencies identified in the published literature, the purposes of this study are:
1. To determine the extent of observer agreement when diagnosing OED;
2. To determine the category specific agreement levels when diagnosing OED;
3. To determine the relative contributions of within- and between-observer disagreement when diagnosing OED;
4. To assess the effect of applying fewer diagnostic categories when diagnosing OED; and,
5. To determine the effect of altered expectation of prior prevalence when diagnosing OED.
Material and Methods

Sample Size
There was little literature available on the sample size necessary to study observer agreement in assessing subjects on a 5 point ordinal scale such as is used when diagnosing OED. Therefore, the sample size used in this study was based on two factors. First, the number of slides was chosen to be convenient for practical reasons. The number of assessments by each pathologist was kept as low as possible while still ensuring adequate sample size. Second, sample size was chosen to be consistent with that used in other observer agreement studies. A table of sample sizes used in various studies is shown in Figure 7. Based on consideration of these two factors, a sample size of 64 cases was chosen. This number is consistent with other studies of observer agreement and convenient for the pathologists involved in the assessments.

Case Selection
Sixty-four histological slides from departmental archives were selected from a sampling frame of 740 cases provided by the Head, Department of Oral Pathology, Faculty of Dentistry, University of Toronto. All cases had a sign-out diagnosis of hyperkeratosis or oral epithelial dysplasia and were assigned a sequential number ordered according to the original date of diagnosis. Potential cases were selected based on computer generated random number lists. Separate lists were used for each category of dysplasia to help ensure equal numbers of cases categorized as normal, mild, moderate, and severe/carcinoma-in-situ at the original sign-out diagnosis. In addition, four cases originally signed out as micro-invasive squamous cell carcinoma were included as study cases. All potential study slides and accompanying referral information were scrutinized by a senior Oral Pathologist with the Department of Oral Pathology and Medicine to ensure they met predetermined inclusion criteria.

Inclusion/Exclusion Criteria
To be selected for the current study, slides had to meet the following criteria: acceptable diagnostic quality; intraoral or labial site; referred by a dentist; and, include referral information on age, sex, and location of the lesion. Exclusion criteria included: inadequate diagnostic quality;
physician referrals; and, cases where adequate referral information was not available or could not be determined. The pathologist selecting the slides did not participate as an examiner in the study. Participating examiners were not informed of the relative number of cases in each diagnostic category.

When a potential case did not meet the inclusion/exclusion criteria, the next case in the sequentially numbered sampling frame was assessed for use as a replacement. The final 64 study cases were originally signed-out as follows: 15 hyperkeratosis with no dysplasia, 15 mild dysplasia, 15 moderate dysplasia, 9 severe dysplasia, 6 carcinoma-in-situ, and 4 micro-invasive squamous cell carcinoma. A computer-generated random number list was used to randomly order the cases. This same random order was utilized for all examiners and examinations.

Examiners/Examinations
Four certified oral pathologists, all affiliated with the Faculty of Dentistry, University of Toronto, were asked to participate in the study. Three agreed to act as observers for the study. For each case, a single archive histological slide stained with haematoxylin and eosin was examined using a light microscope. No special stains or deeper sections were made available. The results of the original signout diagnoses were used as the equivalent of a fourth examiner for comparison and analysis purposes. This methodology was considered appropriate since the signout diagnoses originally represented the diagnostic opinion of a single oral pathologist within the Faculty of Dentistry, University of Toronto.

Study Design
As seen in the study design flowsheet in Figure 8, in the first phase of the study observers were given all 64 cases along with the original referral information on age, sex, smoking status, and anatomical site of the lesion. Observers were asked to complete the examinations within a two week period, assess each slide for the presence and severity of dysplasia, and to record their diagnoses opposite the appropriate lab number in the table provided. No calibration exercises were attempted and dysplasia was graded using the 5-point scale routinely utilized at the Dental
Faculty. Observers were informed that the study cases may include some microinvasive squamous cell carcinomas.

In the second phase of the study, a computer-generated random list of numbers for each diagnostic category was used to separate each category into two roughly equal groups. One group was reexamined with the original referral description and again categorized for the presence and severity of dysplasia by each of the three examining pathologists. These cases were used for the assessment of intraexaminer agreement, i.e., repeat independent observations of an unchanged subject. For the second group, the original referral information was altered prior to repeat observations by participating pathologists. Information on age and smoking status were altered so as to increase or decrease the expected prevalence and/or severity of dysplasia. Observers were again asked to complete the examinations within a two week period, assess each slide for the presence and severity of dysplasia using the same grading scale, and to record their diagnoses opposite the appropriate lab number in the table provided. All assessments were completed within a 12 month period with at least 5 months separating repeat observations of the same cases for each examiner.

Statistical Methods
All random number lists were generated using the Epistat shareware statistical program (83). Data entry was performed by the author using the EpiInfo 5.0 statistical package (84). Data analysis was performed by the author using the SPSS-7.5.1 statistical program (85) and PC-Agree examiner agreement software (68).

Examiner agreement levels were analysed separately for 6 point, 5 point, 3 point, and 2 point scales. The original 6 point ordinal scale consisted of 64 cases distributed in normal, mild, moderate, severe, CIS, and microinvasive categories. The 5 point ordinal scale, consisting of 58 cases distributed in normal, mild, moderate, severe, CIS categories, was created by eliminating all cases categorized as microinvasive carcinoma by at least one observer. The 3 point ordinal scale, consisting of 64 cases distributed in ‘normal’, ‘mild or moderate’, and ‘severe or greater’
categories, was created by collapsing categories in the original data set. The two point dichotomous scale, consisting of 64 cases distributed in 'normal' and 'mild dysplasia or greater' categories, was created by collapsing categories in the original data set.

Inter- and intraobserver agreement was assessed for the 4 different scales by comparison of percent agreement, simple kappa, and weighted kappa using both quadratic and Cicchetti weighting systems. All kappa values were assessed for a statistically significant difference from zero agreement. Simple kappa values for the presence/absence of dysplasia and quadratic weighted kappa values were compared pairwise for statistically significant differences in agreement levels between the different observer pairs. The effect of altering expectation of prior prevalence was assessed for statistical significance with the Wilcoxon signed rank test.

The 'Comparison Standard'
The comparison standard was determined by assessing the 6 point diagnosis provided by the three examiners and the original signout diagnosis for each case (four observations in total). The most frequently repeated (mode) diagnosis for each case was selected as the comparison standard. In 14 cases (21.9%), all 4 observers agreed on the diagnosis. In 31 cases (48.4%), 3 of the 4 observers agreed on the diagnosis. In 10 cases (15.6%), 2 observers agreed on one diagnostic category with the remaining 2 observers each selecting a different diagnostic category. In 8 cases (12.5%), the categorization was evenly split with 2 observers agreeing on one diagnosis while the other two observers agreed on a different diagnosis. In these 8 cases the signout diagnosis was disregarded creating the situation where two of the participating pathologists agreed on a category with the other pathologist disagreeing. This method was used for several reasons. First, the signout diagnosis showed the lowest agreement levels in almost all possible pairwise investigations. Second, disregarding the signout diagnosis resulted in a smaller penalty to the group of participating pathologists than if. Third, the opinion provided with the signout diagnosis may have included information not available during the study such as deeper sections or special stains. Finally, in one case (1.6%), there was no agreement between the 4 different observers. In this case, with diagnoses ranging from normal to severe dysplasia, the signout diagnosis was
disregarded and the middle diagnosis of the three participating pathologists was chosen. In this way the lowest possible penalty would be imposed on the group.

**Observer Blinding**

One consideration in the design of intra-examiner variability studies, is blinding examiners to their previous findings when doing repeat assessment of subjects. Relatively long interval times are necessary to provide adequate blinding to prevent spuriously high self-consistency (39). In this study, at least five months separated the first and second observations for each pathologist thereby minimizing the chance of observer recall bias.

Knowledge of the prior prevalence of a condition has been shown to affect clinical decision making. Prevalence is also known to cause paradoxical behaviour in the kappa statistic used to assess agreement. To avoid problems associated with the effect of prevalence on clinical decision making and the statistics used to assess observer agreement, clinicians were not provided with information regarding the expected distribution of severity of the cases.

**Ethical Consideration**

A full description of the study protocol was submitted to the University of Toronto, Office of Research Services, Ethical Review Committee. Written confirmation was received from the University’s Ethical Review Officer confirming that, due to the negligible risk for involved patients or volunteer pathologists, no formal ethical approval was necessary.
Results

1. General

A total of 448 observations were available for analysis of observer agreement. In addition to the original 64 signout diagnoses, these included 384 new observations made by the three participating pathologists. Subjects included 38 males (59.4%) and 26 females (40.6%) with ages ranging from 20 to 86 years (mean = 54.5 years). Smoking status was distributed as follows: unknown 46.7%, smokers 34.4% and nonsmokers 19.0%. The original location of 1 lesion (1.6%) was not specified, while 23 lesions (35.9%) were from the tongue/floor of mouth, 15 (23.4%) from the gingiva, 6 (9.4%) from the lower lip, 6 (9.4%) from the palate, and 13 (20.3%) from the buccal mucosa.

As with all other studies assessing examiner variability, observer agreement levels displayed in this study were less than perfect. Table 1 shows the relative numbers of cases assigned to each category by the signout diagnosis and the first round of observations. The proportion of cases assigned to each category by each of the three examiners differed slightly from the proportions assigned at the original signout diagnosis. Slightly more cases were assigned to the normal and mild categories, slightly fewer cases to the moderate & severe categories, while the CIS and microinvasive categories remaining relatively unchanged. These differences were not statistically significant (p<0.05) and generally of a magnitude insufficient to create substantial problems with the paradoxical behaviour of kappa.

2. Extent of Observer Agreement

2a. Presence/Absence of Dysplasia

The pairwise percent agreement and simple kappa values for inter- and intraexaminer agreement on assessing the presence or absence of oral epithelial dysplasia are shown in Table 2. Pairwise interexaminer percent agreement is seen to vary from 75 to 84% and the kappa values from 0.42 to 0.58. Combined, the 3 pathologists and signout diagnosis showed 77% agreement on the presence/absence of dysplasia with a group kappa value of 0.51. Intraexaminer agreement varied from 84% to 94% with kappa values ranging from 0.22 to 0.78. All except one of these Kappa
values were significantly different from zero (p<0.001). The intraexaminer kappa for examiner A was not significantly different from zero (p>0.05) despite a reasonably high 84% agreement. The Ppos and Pneg values for this observer were 0.91 and 0.29. As discussed by Cicchetti and Feinstein (63,64), these values show that the lack of significance was due to poor agreement on the absence of dysplasia. Examination of the two-by-two contingency table for examiner A reveals unbalanced marginal totals, the surrogates for prevalence in the absence of an accepted gold standard. The absence of a significant difference from zero agreement can therefore be partly attributed to the first paradox of kappa (63,64). The intraexaminer agreement levels for examiners B and C were both significantly better than that for examiner A (p<0.05).

2b. 6 Point Ordinal Scale
The pairwise percent agreement, simple and weighted kappa values for inter- and intraexaminer agreement on assessing the presence of oral epithelial dysplasia with a 6 point ordinal scale are shown in Table 3. Pairwise interexaminer percent agreement is seen to vary from 47 to 58%, simple kappa values from 0.34 to 0.48, and Cicchetti/quadratic weighted kappa values from 0.59 to 0.70 and 0.66 to 0.86 respectively. Combined, the 3 pathologists and signout diagnosis showed 52% agreement on exact classification with a simple group kappa value of 0.40 and Cicchetti/quadratic weighted group kappa values of 0.63 and 0.75. Intraexaminer agreement varied from 50 to 88%, simple kappa values from 0.36 to 0.84, and Cicchetti/quadratic weighted kappa values ranging from 0.68 to 0.93 and 0.87 to 0.97 respectively. All kappa values are significantly different from zero agreement. The quadratic weighted intraobserver kappa values for examiners B and C were both significantly better than that of examiner A (p<0.01). There was no statistically significant difference between any of the other kappa values.

2c. 5 Point Ordinal Scale
The pairwise percent agreement, simple and weighted kappa values for inter- and intraexaminer agreement on assessing the presence of oral epithelial dysplasia with a 5 point ordinal scale are shown in Table 4. Pairwise interexaminer percent agreement is seen to vary from 45 to 59%, simple kappa values from 0.29 to 0.47, and Cicchetti/quadratic weighted kappa values from 0.55
to 0.67 and 0.68 to 0.83 respectively. Combined, the 3 pathologists and signout diagnosis showed 51% agreement on exact classification with a simple group kappa value of 0.37 and Cicchetti/quadratic weighted group kappa values of 0.59 and 0.74. Intraexaminer agreement varied from 47 to 87%, simple kappa values from 0.30 to 0.83, and Cicchetti/quadratic weighted kappa values ranging from 0.61 to 0.91 and 0.82 to 0.96 respectively. All kappa values are significantly different from zero agreement. The quadratic weighted intraobserver kappa values for examiners B and C were both significantly better than that of examiner A (p<0.001). There was no statistically significant difference between any of the other kappa values.

2d. 3 Point Ordinal Scale
The pairwise percent agreement, simple and weighted kappa values for inter- and intraexaminer agreement on assessing the presence of oral epithelial dysplasia with a 3 point ordinal scale are shown in Table 5. Pairwise interexaminer percent agreement is seen to vary from 67 to 75%, simple kappa values from 0.52 to 0.60, and Cicchetti/quadratic weighted kappa values from 0.58 to 0.71 and 0.60 to 0.81 respectively. Combined, the 3 pathologists and signout diagnosis showed 71% agreement on exact classification with a simple group kappa value of 0.56 and Cicchetti/quadratic weighted group kappa values of 0.61 and 0.68. Intraexaminer agreement varied from 84 to 94%, simple kappa values from 0.74 to 0.90, while Cicchetti/quadratic weighted kappa values ranged from 0.77 to 0.91 and 0.82 to 0.93 respectively. All kappa values are significantly different from zero agreement. There was no statistically significant difference between any of the kappa values.

3. Category Specific Agreement
Table 6 provides category specific agreement results for all four observations combined and for the three participating pathologists separately from the signout diagnosis. Since the differences obtained with three and four are small and identical trends are seen, only the 4 observer results were considered further. When assessing agreement with the 6 point ordinal scale, the four observers in this study were most consistent in applying the lowest (normal) and highest (CIS and microinvasive) categories and less consistent with the middle categories. All observers most
consistently applied the diagnosis of microinvasive squamous cell carcinoma. Using the guidelines developed by Landis & Koch (69), group interobserver agreement levels shows ‘substantial’ agreement beyond chance when classifying microinvasive carcinoma (Ks = 0.685). Group interobserver agreement consistency levels were ‘moderate’ when applying the normal (Ks = 0.510) and carcinoma-in-situ (Ks = 0.543) categories. Group interobserver agreement consistency levels dropped to only ‘fair’ when applying the mild (Ks = 0.261), moderate (Ks = 0.285), and severe (Ks = 0.303) dysplasia categories. A similar trend is seen with the 5 point ordinal scale (which excludes the microinvasive cases).

When collapsed down to a three point ordinal scale, observer simple kappa category specific agreement levels are seen to improve for both of the dysplasia-present ordinal categories (the normal category remains unchanged as it was not involved in the collapsing procedure). Simple kappa values of 0.458 and 0.721 indicate moderate and substantial agreement on applying the ‘mild or moderate’ and ‘severe or greater’ categories, respectively.

Category specific intraobserver agreement results are summarized in Table 7. When assessed for category specific exact intraobserver agreement, all observers showed perfect intraobserver agreement consistency in the application of the microinvasive category. Observers B and C showed better consistency in applying all other diagnostic categories when compared to the group interobserver agreement levels. Agreement in most categories was in the ‘almost perfect’ range. Observer A’s consistency in exact category agreement was significantly worse (P<0.01) than those of observers B and C.

As summarized in Table 8, when assessed for conformity with the modal diagnosis ‘comparison standard,’ the accuracy for all four observers was in the ‘almost perfect’ category for application of the microinvasive carcinoma and carcinoma-in-situ categories. ‘Substantial’ agreement was generally seen in the accurate application of normal and severe categories, while accuracy in categorization of mild and moderate cases ranged from ‘fair’ to ‘almost perfect.’
4. Conformity with Modal Diagnosis

As seen in Tables 2 to 5, when the modal diagnosis is used as a 'comparison standard' to assess conformity, observer percent agreement levels range from 86% to 89% and Ks from 0.65 to 0.76 (substantial) for the presence or absence of dysplasia. With the 6 point scale percent agreement ranges from 64% to 80% and quadratic Kw from 0.81 to 0.92 (almost perfect). The 3 point scale shows percent agreement levels that range from 78% to 86% and quadratic Kw from 0.74 to 0.88 (substantial to almost perfect). All observers showed significantly greater than zero agreement. There were no statistically significant (p>0.05) differences in the agreement levels shown by the four observer groups.

5. Effect of Altered Expectation of Prior Prevalence

As previously described, of the 64 cases initially provided to the cooperating oral pathologists with the original referral information, one half were reexamined with the referral information unchanged. For the remaining 32 cases, prior to the second set of observations, the information regarding age and smoking status was altered to change the observer's expectation of prior prevalence. As seen in Table 9, for the 15 cases where age was reported to be higher and smoking status increased from the initial examination, 6.7%, 62.2%, and 31.1% were judged as being less severe, equal, or more severe than at the first examination. The figures resulting from the 17 cases where the age was reported to be lower and smoking status decreased from the initial examination were 29.4%, 58.8%, and 11.8% respectively.

These differences were assessed for statistical significance with the Wilcoxon Signed Rank Test. When the expectation of prior prevalence was increased, a significant number of increased severity diagnoses were given by the group (p<0.01), observer A (p<0.05) and observer B (p<0.05) but not observer C (p>0.05). When the expectation of prior prevalence remained the same, there was no significant change (p>0.05) in diagnosed severity for the group, observer A or observer C while observer B increased diagnostic severity in a small but significant (p>0.05) number of cases. This was the only instance where a significant change occurred that was contrary to what was anticipated. When the expectation of prior prevalence was decreased, a
significant number of decreased severity diagnoses were given by observer C \( (p<0.05) \) but not observers A or B \( (p>0.05) \). Although a definite trend to decrease the severity of dysplasia was seen for the group when expectation of prior prevalence was decreased, the differences did not quite reach statistical significance \( (0.10>p>0.05) \).
Discussion

This study selected equal numbers of histological slides, based on the original signout diagnosis, covering the full range of categories used in the assessment of oral epithelial dysplasia. Three participating oral pathologists assessed each slide two times to allow evaluation of the degree of observer agreement consistency and the effect of altering the expectation of prior prevalence on the diagnosis of OED. By determining the most common diagnosis given for each case (modal diagnosis), the accuracy of each pathologist’s diagnosis was investigated by looking at the level of conformity with this standard.

Although the Cicchetti kappa weighting system was initially developed for use on mixed dichotomous-ordinal classification systems such as that used for categorizing the presence and severity of OED, the Kw values obtained using the quadratic weighting system were relied upon to most accurately show observer agreement levels. The quadratic weights were selected since the patient management differences resulting from a single category discrepancy around the normal-mild dysplasia categories are minimal. With mild dysplasia, most clinicians tend to recommend noninvasive preventive treatments such as smoking cessation and alcohol risk reduction counselling. These interventions are also appropriate for patients without dysplastic oral lesions. Therefore, the additional penalty imposed by the Cicchetti weighting system on discrepancies between the presence and absence of dysplasia are unjustified.

This study shows that the extent of interobserver agreement when diagnosing oral epithelial dysplasia on a 6 or 5 point scale is in the substantial to almost perfect range (0.66 to 0.86 and 0.68 to 0.83 respectively) when assessed with the appropriate quadratic weighted kappa statistic. These excellent agreement levels were comforting to find since they have such important patient treatment implications. The participating pathologists in this study generally showed simple kappa values equal to or slightly better than those previously reported in the oral epithelial dysplasia literature (22,55). Whether these differences in simple kappa values are significant or not could not be assessed. However, they are of a magnitude unlikely to be statistically significant.
This is the first OED observer agreement study to report agreement levels with the appropriate weighted kappa statistic, thus no comparison with previous studies is possible. However, in one study of observer variability in assessing oral premalignant lesions, Karabulut et al. (22) provided the raw data for one pair of observers. All observer pairs in the current study showed better observer agreement levels than those displayed by the Karabulut study. This difference was only statistically significant when this study’s observers A and C pairwise agreement was compared with the 1995 study pair (p<0.01).

Category specific agreement levels in diagnosis of OED have not been previously published in the scientific literature. This study found that the extent of category specific agreement, assessed by simple kappa, for three participating pathologists diagnosing oral epithelial dysplasia on a 6 or 5 point ordinal scale was: substantial for microinvasive carcinoma (0.72); moderate to substantial for CIS (0.61 to 0.65); moderate for severe dysplasia (0.49 to 0.51); fair for moderate dysplasia (0.31 to 0.32); slight for mild dysplasia (0.14 to 0.17); and, moderate for the absence of dysplasia (0.45). This trend of improved agreement at both ends of the scale and best agreement with the most severe cases is similar to that reported for cervical intraepithelial neoplasia (24,53,77).

Having observers agree best with the most severe cases is reassuring since these diagnostic decisions result in the most significant patient treatment implications. Higher levels of observer agreement seen with the more advanced cases show that relatively few cases are either over classified, resulting in unnecessary patient surgery, or under-called, resulting in a lesion with high malignant potential being left in the patient’s mouth. Moderate agreement above chance on the absence of dysplasia shows that the majority of normal cases will be diagnosed as such. As shown in Table 11, when disagreement on the presence of dysplasia does occur, there is unlikely to be more than one category of discrepancy.

Since the University of Toronto’s practice is to recommend treating mild and moderate dysplasia cases with preventive interventions, the health and future treatment implications for these patients is minimal. These findings support the current recommendation that lesions with severe dysplasia
or greater be surgically treated. It is with these lesions that the greatest confidence in observer accuracy is appropriate.

When the results of both the inter- and intraobserver analyses are considered, this study shows that intraobserver agreement (0.87 to 0.97) levels surpass those of interobserver agreement (0.66-0.86). Although there is some random error present in the diagnosis of OED (intraobserver agreement slightly less than perfect), it is not as important a consideration as the systematic error (interobserver agreement lower than intraobserver agreement). This observer bias arises from the use of slightly different diagnostic thresholds and results in slightly different proportions of cases assigned to each diagnostic category. This type of systematic error can often be explained by differences in training (position of thresholds) as well as different opinions on which of the many criteria are most important in the categorization of dysplasia (22,55,72). Although this study’s interobserver agreement levels are relatively good, because observer bias is present, the observed systematic variation might decrease if calibration exercises were instigated. On the other hand, as discussed previously, random error can prove to be very difficult to overcome. In the case of oral epithelial dysplasia, this small residual random error may reflect a combination of imperfect understanding of the disease process and, as discussed by Morris (49), the random categorization of cases which fall between the centre points of adjacent categories.

The intra- and interobserver agreement levels for one of the participating pathologists warrant further discussion. When assessed for interobserver agreement, observer A showed the best pairwise interexaminer agreement (0.86) and had the best consistency with the signout diagnosis (0.80) and conformity with the modal diagnosis (0.92). When compared with the modal diagnosis for the presence/absence of dysplasia, observer A showed a simple kappa of 0.73. These high interobserver agreement and conformity levels suggest that high confidence for accuracy is appropriate for observer A. Unexpectedly, however, observer A’s intraobserver agreement levels were the lowest of the three participating pathologists (0.87 vs 0.97/0.97) and for the presence or absence of dysplasia were not significantly better than zero agreement (0.22 vs 0.76/0.78 with p>0.05). These findings are contrary to the usual situation where intraobserver agreement levels
surpass interobserver agreement levels. A possible explanation for this finding lies in the substantial time and effort required by participating pathologists when viewing this large series of histologic slides. The time and effort expended in the first round of observations may have not have been duplicated in the second round of observations with less time and importance devoted to assessing the study materials on repeat examination. Observer fatigue, as listed in the ‘sources of variation’ section of this manuscript, may have been an important component of observer A’s intraobserver variation. Why this may have affected observer A more than observers B or C is not known.

This study shows that decisions regarding the relative merit of utilizing fewer diagnostic categories depend on the statistic being utilized for the analysis. As shown in Table 10, if percent agreement is used as the statistic for analysis, observer agreement levels improve substantially (10% to 25% increase) when the diagnosis of OED is reduced to 3 categories which correspond to the commonly recommended treatment thresholds (compare Tables 3 and 5). Fewer categories thus appear to be advantageous. Similarly, if simple kappa is used for the analysis, observer agreement levels improve (0.05 to 0.20 increase) when a 3 category system is used. Again, fewer categories appear to improve the problem of poor observer agreement levels. However, if weighted kappa is used for the analysis, observer agreement levels drop slightly (0.03 to 0.09 decrease). Thus, when using weighted kappa, the statistic of choice for analysing observer agreement with ordinal data, there is no advantage gained by reducing the number of diagnostic categories.

Support for use of the full 5 item diagnostic scale comes from considering the extent of disagreement when it does occur. Table 11 shows the number of categories of discrepancy seen when comparing observations made by the participating pathologists with the modal diagnosis. Perfect category specific agreement occurred in 75% of the 192 observations performed by the three participating pathologists. Of the 48 (25%) observations not showing perfect agreement, only 8 (4.2%) showed a discrepancy of more than one category. Based on these findings, continued use of the full ordinal scale of dysplasia categorization is recommended. Previously
published results on this subject are not available for comparison.

It is important to note that the three point ordinal scale utilized in this study was created after the diagnosis by collapsing categories in the full six point scale. It is possible that different results would be achieved if pathologists directly allocated the cases to a 3 point ordinal scale. However, for the purposes of this study, it was assumed that collapsing categories gave results no different from those likely to be obtained with the direct allocation of cases. This assumption remains to be proven and is a subject for future research on oral epithelial dysplasia.

This experiment showed that information which alters the expectation of prior prevalence often influences the subsequent diagnosis of dysplasia. Patient age and smoking status are two factors known to be associated with increasing prevalence and severity of oral epithelial dysplasia. If a patient with a suspicious intraoral lesion is an elderly smoker, the chance of dysplasia being present is higher than for a younger non-smoker. In other words, even before applying any diagnostic tests (evaluating the presence of the various histological features listed in Figure 1) the clinician knows that the prevalence of dysplasia in this population (elderly smokers) is higher than in the general population and is appropriately more suspicious of the presence of dysplasia (47). As shown in Table 9, when given patient information that either increases or decreases expectation of prior prevalence, pathologists are more likely to rate the dysplasia higher or lower, respectively. This finding is consistent with what has been reported in the radiology literature where clinical history (40,43) and information about the prevalence of disease (42) has been shown to influence diagnostic decisions. No other studies have looked at the effect of expectation of prior prevalence on the diagnosis of oral epithelial dysplasia.

The effect of expectation of prior prevalence has important clinical implications. It is common for the University of Toronto oral pathology service to receive biopsy specimens accompanied with little or no patient information. When key patient information (i.e., age, smoking status, alcohol status, lesion topography) is missing, it cannot impart an effect on expectation of prior prevalence. Similarly, inaccurate information provided by the referring clinician could have a detrimental
effect on the accuracy of the subsequent diagnosis. Therefore, referring clinicians should be advised to include the pertinent patient history and a complete description of the oral lesion with every tissue specimen submitted.

In final summary, the results and implications of this study are:

1. Using weighted kappa, there is substantial interobserver agreement (Group Kw = 0.74) and almost perfect intraobserver agreement (Kw = 0.87 to 0.97) in the diagnosis of OED amongst the participating pathologists at the Faculty of Dentistry, University of Toronto.
2. The three pathologists show the best agreement when diagnosing the most severe cases of dysplasia (K = 0.30-0.57), an intermediate level of agreement when diagnosing the presence/absence of dysplasia (K = 0.51), and the lowest agreement when diagnosing the less severe dysplastic cases (K = 0.24-0.26).
3. Bias is the most important source of disagreement since most variation was between-observer as evidenced by intraobserver agreement levels which surpass interobserver agreement levels. Consensus and calibration exercises may, therefore, be appropriate.
4. Reducing the number of diagnostic categories does not improve observer agreement when analysed with the appropriate statistic (Kw). Since the best agreement is seen with higher severity lesions with their important treatment implications, the full 5 point ordinal scale should continue to be used when assessing OED.
5. Information available at the time of histopathological evaluation can have a significant influence on the diagnosis of oral epithelial dysplasia. When given a patient history which suggests a higher prevalence of dysplasia, observers are significantly more likely to grade the prevalence or severity higher than if given a patient history which suggests a lower prevalence. Referring clinicians should, therefore, be advised to include an accurate patient history and lesion description with every tissue specimen submitted.
Table 1 - Classification of lesions by signout diagnosis and three independent observers.

<table>
<thead>
<tr>
<th>Cross tabulation</th>
<th>None</th>
<th>Mild</th>
<th>Moderate</th>
<th>Severe</th>
<th>CIS</th>
<th>Micro invasive</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full 6 point scale</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>17</td>
<td>17</td>
<td>9</td>
<td>11</td>
<td>6</td>
<td>4</td>
<td>64</td>
</tr>
<tr>
<td>Mild</td>
<td>26.6</td>
<td>26.6</td>
<td>14.1</td>
<td>17.2</td>
<td>9.4</td>
<td>6.3</td>
<td>100</td>
</tr>
<tr>
<td>Moderate</td>
<td>15</td>
<td>20</td>
<td>15</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>64</td>
</tr>
<tr>
<td>Severe</td>
<td>23.4</td>
<td>31.3</td>
<td>23.4</td>
<td>9.4</td>
<td>7.8</td>
<td>4.7</td>
<td>100</td>
</tr>
<tr>
<td>CIS</td>
<td>4.7</td>
<td>14</td>
<td>15.5</td>
<td>7.8</td>
<td>14.1</td>
<td>9.4</td>
<td>100</td>
</tr>
<tr>
<td>Micro invasive</td>
<td>9.4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>9.4</td>
<td>6.3</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>26.6</td>
<td>26.6</td>
<td>14.1</td>
<td>17.2</td>
<td>9.4</td>
<td>6.3</td>
<td>100</td>
</tr>
</tbody>
</table>

Examiner A: Count 17, 17, 9, 11, 6, 4, 64
Row % 26.6, 26.6, 14.1, 17.2, 9.4, 6.3, 100

Examiner B: Count 15, 20, 15, 6, 5, 3, 64
Row % 23.4, 31.3, 23.4, 9.4, 7.8, 4.7, 100

Examiner C: Count 25, 9, 10, 5, 9, 6, 64
Row % 39.1, 14.1, 15.6, 7.8, 14.1, 9.4, 100

Sign out: Count 15, 15, 15, 9, 6, 4, 64
Row % 23.4, 23.4, 23.4, 14.1, 9.4, 6.3, 100

Total: Count 72, 61, 49, 31, 26, 17, 256
Row % 28.1, 23.8, 19.1, 12.1, 10.2, 6.6, 100
Table 2 - Observer agreement on presence/absence of dysplasia

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Signout</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>% - 84%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ks - 0.22*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>% - 78%</td>
<td>% - 94%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ks - 0.42</td>
<td>Ks - 0.76 **</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>% - 78%</td>
<td>% - 75%</td>
<td>% - 91%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ks - 0.51</td>
<td>Ks - 0.43</td>
<td>Ks - 0.78 **</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Signout</td>
<td>% - 84%</td>
<td>% - 84%</td>
<td>% - 81%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ks - 0.58</td>
<td>Ks - 0.56</td>
<td>Ks - 0.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mode</td>
<td>% - 89%</td>
<td>% - 86%</td>
<td>% - 89%</td>
<td>% - 89%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ks - 0.73</td>
<td>Ks - 0.65</td>
<td>Ks - 0.76</td>
<td>Ks - 0.73</td>
<td></td>
</tr>
</tbody>
</table>

Group % agreement = 77%
Group kappa = 0.51; 95% CI = [0.42 - 0.61]

- **Consistency - Intraobserver agreement**
- **Consistency - Interobserver pairwise agreement**
- **Conformity - With 'Comparison Standard' = Mode Diagnosis**

* Kappa value is not significantly different from zero (p>0.05). Ppos = 0.91; Pneg = 0.29. All other Kappa values are significantly different from zero agreement.
** Intraobserver agreement levels for examiners B and C are significantly better (p<0.05) than that of observer A. There is no statistically significant difference between any of the other kappa values.
Table 3 - Observer agreement with 6 point ordinal scale

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Signout</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>% - 50%</td>
<td>% - 88%</td>
<td>% - 52%</td>
<td>% - 47%</td>
<td>% - 69%</td>
</tr>
<tr>
<td></td>
<td>Ks - 0.36</td>
<td>Ks - 0.37</td>
<td>Ks - 0.40</td>
<td>Ks - 0.34</td>
<td>Ks - 0.61</td>
</tr>
<tr>
<td></td>
<td>Kw - 0.68/0.87</td>
<td>Kw - 0.60/0.72</td>
<td>Kw - 0.70/0.86</td>
<td>Kw - 0.63/0.80</td>
<td>Kw - 0.80/0.92</td>
</tr>
<tr>
<td>B</td>
<td>% - 50%</td>
<td>% - 88%</td>
<td>% - 52%</td>
<td>% - 47%</td>
<td>% - 69%</td>
</tr>
<tr>
<td></td>
<td>Ks - 0.36</td>
<td>Ks - 0.37</td>
<td>Ks - 0.40</td>
<td>Ks - 0.34</td>
<td>Ks - 0.61</td>
</tr>
<tr>
<td></td>
<td>Kw - 0.68/0.87</td>
<td>Kw - 0.60/0.72</td>
<td>Kw - 0.70/0.86</td>
<td>Kw - 0.63/0.80</td>
<td>Kw - 0.80/0.92</td>
</tr>
<tr>
<td>C</td>
<td>% - 52%</td>
<td>% - 88%</td>
<td>% - 52%</td>
<td>% - 48%</td>
<td>% - 77%</td>
</tr>
<tr>
<td></td>
<td>Ks - 0.40</td>
<td>Ks - 0.47</td>
<td>Ks - 0.40</td>
<td>Ks - 0.34</td>
<td>Ks - 0.61</td>
</tr>
<tr>
<td></td>
<td>Kw - 0.70/0.86</td>
<td>Kw - 0.66/0.74</td>
<td>Kw - 0.66/0.74</td>
<td>Kw - 0.60/0.66</td>
<td>Kw - 0.80/0.81</td>
</tr>
<tr>
<td>Signout</td>
<td>% - 47%</td>
<td>% - 58%</td>
<td>% - 48%</td>
<td>% - 48%</td>
<td>% - 64%</td>
</tr>
<tr>
<td></td>
<td>Ks - 0.34</td>
<td>Ks - 0.47</td>
<td>Ks - 0.36</td>
<td>Ks - 0.36</td>
<td>Ks - 0.55</td>
</tr>
<tr>
<td></td>
<td>Kw - 0.63/0.80</td>
<td>Kw - 0.60/0.66</td>
<td>Kw - 0.60/0.66</td>
<td>Kw - 0.59/0.74</td>
<td>Kw - 0.83/0.91</td>
</tr>
<tr>
<td>Mode</td>
<td>% - 69%</td>
<td>% - 80%</td>
<td>% - 77%</td>
<td>% - 77%</td>
<td>% - 64%</td>
</tr>
<tr>
<td></td>
<td>Ks - 0.61</td>
<td>Ks - 0.75</td>
<td>Ks - 0.70</td>
<td>Ks - 0.70</td>
<td>Ks - 0.55</td>
</tr>
<tr>
<td></td>
<td>Kw - 0.80/0.92</td>
<td>Kw - 0.80/0.81</td>
<td>Kw - 0.80/0.81</td>
<td>Kw - 0.83/0.91</td>
<td>Kw - 0.71/0.83</td>
</tr>
</tbody>
</table>

Kw = Cicchetti/Quadratic weights

Group % agreement = 52%
Simple group kappa = 0.40; 95% CI = [0.35 - 0.45]
Quadratic weighted group kappa = 0.75; 95% CI = [0.65 - 0.85]

- All kappa values are significantly different from zero agreement.
* Intraobserver kappa values for observers B and C are both significantly better than that of observer A (p<0.01). There is no statistically significant difference between any of the other kappa values.
Table 4 - *Pairwise Observer agreement with 5 point ordinal scale*

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Signout</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>% - 47%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ks - 0.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kw - 0.61/0.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>% - 48%</td>
<td>% - 87%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ks - 0.33</td>
<td>Ks - 0.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kw - 0.57/0.72</td>
<td>Kw - 0.91/0.96*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>% - 50%</td>
<td>% - 59%</td>
<td>% - 76%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ks - 0.35</td>
<td>Ks - 0.47</td>
<td>Ks - 0.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kw - 0.65/0.83</td>
<td>Kw - 0.67/0.79</td>
<td>Kw - 0.86/0.95*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Signout</td>
<td>% - 45%</td>
<td>% - 57%</td>
<td>% - 48%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ks - 0.29</td>
<td>Ks - 0.44</td>
<td>Ks - 0.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kw - 0.56/0.75</td>
<td>Kw - 0.57/0.68</td>
<td>Kw - 0.55/0.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mode</td>
<td>% - 67%</td>
<td>% - 78%</td>
<td>% - 78%</td>
<td>% - 60%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ks - 0.58</td>
<td>Ks - 0.71</td>
<td>Ks - 0.70</td>
<td>Ks - 0.49</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kw - 0.76/0.88</td>
<td>Kw - 0.77/0.80</td>
<td>Kw - 0.84/0.92</td>
<td>Kw - 0.64/0.76</td>
<td></td>
</tr>
</tbody>
</table>

Kw = Cicchetti/Quadratic weights

Group % agreement = 51%
Simple group kappa = 0.37; 95% CI = [0.32 - 0.42]
Quadratic weighted group kappa = 0.74; 95% CI = [0.64 - 0.85]

- All kappa values are significantly different from zero agreement.
- Intraobserver kappa values for observers B and C are both significantly better than that of observer A (p<0.001). There is no statistically significant difference between any of other kappa values.
Table 5 - Observer agreement with 3 point ordinal scale

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Signout</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% - 84%</td>
<td>% - 70%</td>
<td>% - 72%</td>
<td>% - 72%</td>
<td>% - 86%</td>
</tr>
<tr>
<td></td>
<td>Ks - 0.74</td>
<td>Ks - 0.55</td>
<td>Ks - 0.60</td>
<td>Ks - 0.57</td>
<td>Ks - 0.79</td>
</tr>
<tr>
<td></td>
<td>Kw - 0.77/0.82</td>
<td>Kw - 0.61/0.66</td>
<td>Kw - 0.71/0.81</td>
<td>Kw - 0.65/0.75</td>
<td>Kw - 0.83/0.88</td>
</tr>
<tr>
<td>B</td>
<td>% - 94%</td>
<td>% - 67%</td>
<td>% - 67%</td>
<td>% - 75%</td>
<td>% - 81%</td>
</tr>
<tr>
<td></td>
<td>Ks - 0.90</td>
<td>Ks - 0.53</td>
<td>Ks - 0.60</td>
<td>Ks - 0.60</td>
<td>Ks - 0.71</td>
</tr>
<tr>
<td></td>
<td>Kw - 0.91/0.93</td>
<td>Kw - 0.60/0.65</td>
<td>Kw - 0.60/0.65</td>
<td>Kw - 0.60/0.60</td>
<td>Kw - 0.72/0.74</td>
</tr>
<tr>
<td>C</td>
<td>% - 91%</td>
<td>% - 67%</td>
<td>% - 67%</td>
<td>% - 52%</td>
<td>% - 84%</td>
</tr>
<tr>
<td></td>
<td>Ks - 0.86</td>
<td>Ks - 0.52</td>
<td>Ks - 0.57</td>
<td>Ks - 0.77</td>
<td>Ks - 0.77</td>
</tr>
<tr>
<td></td>
<td>Kw - 0.90/0.93</td>
<td>Kw - 0.58/0.66</td>
<td>Kw - 0.60/0.60</td>
<td>Kw - 0.82/0.88</td>
<td>Kw - 0.71/0.77</td>
</tr>
</tbody>
</table>

Kw = Cicchetti/Quadratic weights

Group % agreement = 71%
Simple group kappa = 0.56; 95% CI = [0.49 - 0.63]
Quadratic weighted group kappa = 0.68; 95% CI = [0.58 - 0.78]

- Consistency - Intraobserver agreement
- Consistency - Interobserver pairwise agreement
- Conformity - With 'Comparison Standard' = Mode Diagnosis

- All kappa values are significantly different from zero agreement.
- There is no statistically significant difference between any of the kappa values.
Table 6 - Category specific group interobserver agreement with 3 observers and 3 observers plus signout diagnosis.

<table>
<thead>
<tr>
<th>Category</th>
<th>Ks - 3 Observers</th>
<th>Ks - 4 Observers</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.451</td>
<td>0.51</td>
</tr>
<tr>
<td>Mild</td>
<td>0.171</td>
<td>0.261</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.321</td>
<td>0.285</td>
</tr>
<tr>
<td>Severe</td>
<td>0.487</td>
<td>0.303</td>
</tr>
<tr>
<td>CIS</td>
<td>0.609</td>
<td>0.543</td>
</tr>
<tr>
<td>Microinvasive</td>
<td>0.752</td>
<td>0.685</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Ks - 3 Observers</th>
<th>Ks - 4 Observers</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.447</td>
<td>0.506</td>
</tr>
<tr>
<td>Mild</td>
<td>0.143</td>
<td>0.236</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.311</td>
<td>0.259</td>
</tr>
<tr>
<td>Severe</td>
<td>0.513</td>
<td>0.304</td>
</tr>
<tr>
<td>CIS</td>
<td>0.646</td>
<td>0.567</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Ks - 3 Observers</th>
<th>Ks - 4 Observers</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.451</td>
<td>0.51</td>
</tr>
<tr>
<td>Mild or Moderate</td>
<td>0.443</td>
<td>0.458</td>
</tr>
<tr>
<td>Severe or Greater</td>
<td>0.796</td>
<td>0.721</td>
</tr>
</tbody>
</table>
Table 7 - Overall and category specific intraobserver agreement

<table>
<thead>
<tr>
<th></th>
<th>Overall Ks/Kw*</th>
<th>Normal</th>
<th>Mild</th>
<th>Moderate</th>
<th>Severe</th>
<th>CIS</th>
<th>Microinvasive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observer A</strong></td>
<td>0.36/0.87</td>
<td>0.198</td>
<td>0.307</td>
<td>0.354</td>
<td>0.5</td>
<td>-0.103</td>
<td>1</td>
</tr>
<tr>
<td><strong>Observer B</strong></td>
<td>0.84/0.97</td>
<td>0.763</td>
<td>0.845</td>
<td>1</td>
<td>0.632</td>
<td>0.795</td>
<td>1</td>
</tr>
<tr>
<td><strong>Observer C</strong></td>
<td>0.73/0.97</td>
<td>0.775</td>
<td>0.451</td>
<td>0.714</td>
<td>0.632</td>
<td>0.817</td>
<td>1</td>
</tr>
</tbody>
</table>

* Weighted kappa reported using quadratic weights. All others reported using simple kappa.
Table 8 - *Overall and category specific conformity with modal diagnosis*

<table>
<thead>
<tr>
<th></th>
<th>Overall*</th>
<th>Normal</th>
<th>Mild</th>
<th>Moderate</th>
<th>Severe</th>
<th>CIS</th>
<th>Microinvasive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer A</td>
<td>0.9164</td>
<td>0.734</td>
<td>0.42</td>
<td>0.341</td>
<td>0.691</td>
<td>0.743</td>
<td>1</td>
</tr>
<tr>
<td>Observer B</td>
<td>0.8056</td>
<td>0.646</td>
<td>0.667</td>
<td>0.909</td>
<td>0.679</td>
<td>0.816</td>
<td>0.849</td>
</tr>
<tr>
<td>Observer C</td>
<td>0.9106</td>
<td>0.76</td>
<td>0.487</td>
<td>0.629</td>
<td>0.743</td>
<td>0.857</td>
<td>0.783</td>
</tr>
<tr>
<td>Signout</td>
<td>0.8275</td>
<td>0.725</td>
<td>0.578</td>
<td>0.451</td>
<td>0.254</td>
<td>0.519</td>
<td>0.849</td>
</tr>
</tbody>
</table>

*Overall agreement reported using weighted kappa and quadratic weights. All others reported using simple kappa.*

- All Kappa values are significantly different from zero agreement.
- There is no statistically significant difference between any of the Kappa values.
Table 9 - Effect of prior prevalence expectation on diagnosis of dysplasia

<table>
<thead>
<tr>
<th>Expectation of Prior Prevalence - Unchanged</th>
<th>Observer A</th>
<th>Observer B</th>
<th>Observer C</th>
<th>Group (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>13 (13.5%)</td>
</tr>
<tr>
<td>Same</td>
<td>16</td>
<td>28</td>
<td>25</td>
<td>69 (71.9%)</td>
</tr>
<tr>
<td>Decreased</td>
<td>8</td>
<td>0</td>
<td>6</td>
<td>14 (14.6%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expectation of Prior Prevalence - Increased</th>
<th>Observer A</th>
<th>Observer B</th>
<th>Observer C</th>
<th>Group (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>14 (31.1%)</td>
</tr>
<tr>
<td>Same</td>
<td>7</td>
<td>10</td>
<td>11</td>
<td>28 (62.2%)</td>
</tr>
<tr>
<td>Decreased</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3 (6.7%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expectation of Prior Prevalence - Decreased</th>
<th>Observer A</th>
<th>Observer B</th>
<th>Observer C</th>
<th>Group (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>6 (11.8%)</td>
</tr>
<tr>
<td>Same</td>
<td>7</td>
<td>12</td>
<td>11</td>
<td>30 (58.8%)</td>
</tr>
<tr>
<td>Decreased</td>
<td>7</td>
<td>2</td>
<td>6</td>
<td>15 (29.4%)</td>
</tr>
</tbody>
</table>

Wilcoxon Signed Ranks Test

<table>
<thead>
<tr>
<th>Expectation of Prior Prevalence</th>
<th>Ranks Neg/Pos</th>
<th>Mean Ranks (Neg/Pos)</th>
<th>Ranks Sum (Neg/Pos)</th>
<th>Z-Score</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increased</td>
<td>3/14</td>
<td>8.50/9.11</td>
<td>25.5/127.5</td>
<td>-2.683</td>
<td>0.007</td>
</tr>
<tr>
<td>Unchanged</td>
<td>14/13</td>
<td>14.00/14.00</td>
<td>196.0/182.0</td>
<td>-0.102</td>
<td>0.847</td>
</tr>
<tr>
<td>Decreased</td>
<td>15/6</td>
<td>10.70/11.75</td>
<td>160.5/70.5</td>
<td>-1.719</td>
<td>0.086</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observer A</th>
<th>Ranks Neg/Pos</th>
<th>Mean Ranks (Neg/Pos)</th>
<th>Ranks Sum (Neg/Pos)</th>
<th>Z-Score</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased</td>
<td>1/7</td>
<td>4.50/4.50</td>
<td>4.5/31.5</td>
<td>-2.121</td>
<td>0.034</td>
</tr>
<tr>
<td>Unchanged</td>
<td>8/8</td>
<td>8.50/8.50</td>
<td>68.0/68.0</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Decreased</td>
<td>7/3</td>
<td>5.00/6.67</td>
<td>35.0/20.0</td>
<td>-0.832</td>
<td>0.405</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observer B</th>
<th>Ranks Neg/Pos</th>
<th>Mean Ranks (Neg/Pos)</th>
<th>Ranks Sum (Neg/Pos)</th>
<th>Z-Score</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased</td>
<td>0/5</td>
<td>0.00/3.00</td>
<td>0.0/15.0</td>
<td>-2.121</td>
<td>0.034</td>
</tr>
<tr>
<td>Unchanged</td>
<td>0/4</td>
<td>0.00/2.50</td>
<td>0.0/10.0</td>
<td>-2.00</td>
<td>0.046</td>
</tr>
<tr>
<td>Decreased</td>
<td>2/3</td>
<td>3.00/3.00</td>
<td>6.0/9.0</td>
<td>-0.447</td>
<td>0.655</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observer C</th>
<th>Ranks Neg/Pos</th>
<th>Mean Ranks (Neg/Pos)</th>
<th>Ranks Sum (Neg/Pos)</th>
<th>Z-Score</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased</td>
<td>2/2</td>
<td>2.50/2.50</td>
<td>5.0/5.0</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Unchanged</td>
<td>6/1</td>
<td>4.00/4.00</td>
<td>24.0/4.0</td>
<td>-1.890</td>
<td>0.059</td>
</tr>
<tr>
<td>Decreased</td>
<td>6/0</td>
<td>3.50/0.00</td>
<td>21.0/0.0</td>
<td>-2.333</td>
<td>0.020</td>
</tr>
</tbody>
</table>
Table 10 - Comparison of group and category specific interobserver agreement with 6 and 3 point ordinal scales

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>6 Point</th>
<th>3 Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>% agreement</td>
<td></td>
<td>52%</td>
<td>71%</td>
</tr>
<tr>
<td>Ks</td>
<td></td>
<td>0.4</td>
<td>0.56</td>
</tr>
<tr>
<td>Kw</td>
<td></td>
<td>0.75</td>
<td>0.68</td>
</tr>
</tbody>
</table>

**Category Specific**

<table>
<thead>
<tr>
<th>Category</th>
<th>6 Point</th>
<th>3 Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>Mild</td>
<td>0.261</td>
<td>0.458</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.285</td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td>0.303</td>
<td></td>
</tr>
<tr>
<td>CIS</td>
<td>0.543</td>
<td>0.721</td>
</tr>
<tr>
<td>Microinvasive</td>
<td>0.685</td>
<td></td>
</tr>
</tbody>
</table>
Table 11 - *Number of categories of discrepancy from modal diagnosis*

### 6 Point Ordinal Scale

<table>
<thead>
<tr>
<th>Examiner</th>
<th>Agreement</th>
<th>1 Category</th>
<th>2 Categories</th>
<th>3 Categories</th>
<th>4 Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>44</td>
<td>19</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>51</td>
<td>10</td>
<td>1</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>49</td>
<td>11</td>
<td>3</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

### 3 Point Ordinal Scale

<table>
<thead>
<tr>
<th>Examiner</th>
<th>Agreement</th>
<th>1 Category</th>
<th>2 Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>55</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>53</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>54</td>
<td>10</td>
<td>-</td>
</tr>
</tbody>
</table>


Figure 1. *Epithelial changes which may be seen in epithelial dysplasia*

**Disorderly Maturation:**
- irregular hyperplasia and/or atrophy
- keratosis/parakeratosis
- drop-shaped rete processes
- irregular stratification and disturbed cell polarity
- low level keratinization in single or small cell groups
- reduced epithelial cell cohesion
- cell pleomorphism

**Disturbed Cell Proliferation:**
- loss of basal cell polarity
- basal cell hyperplasia
- increased nuclear-cytoplasmic ratio
- enlarged nucleoli
- nuclear hyperchromatism
- high-level mitoses
- anisonucleosis
- abnormal ('bizarre') mitoses

Figure 2. *The etiology of clinical disagreement*

**The Examiner**
1. Biologic variation in the acuity of clinician senses
2. Tendency to record inference rather than evidence
3. Ensnarement by diagnostic classification schemes
4. Entrapment by prior expectation
5. Simple incompetency

**The Examined**
1. Biologic variation in the system being examined
2. Effects of illness and medications
3. Memory and rumination
4. Toss-ups

**The Examination**
1. Disruptive environments
2. Disruptive interactions between examiners and examined
3. Incorrect function or use of diagnostic tools

Figure 3. Example of two examiners showing perfect correlation with no concordance.
Figure 4. Calculating Agreement with the 2x2 Contingency Table

<table>
<thead>
<tr>
<th>Observation 2</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>No</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Observation 1

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>a</td>
<td>R1</td>
</tr>
<tr>
<td>No</td>
<td>c</td>
<td>R2</td>
</tr>
</tbody>
</table>

where  
R = row  
C = column  
T = total
Figure 5. *The dependence of kappa on disease prevalence for a fixed level of observer agreement.*

Figure 6. *Quantitative significance of Kappa*

<table>
<thead>
<tr>
<th>Value of Ks</th>
<th>Strength of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0</td>
<td>Poor</td>
</tr>
<tr>
<td>0.0 - .20</td>
<td>Slight</td>
</tr>
<tr>
<td>.21 - .40</td>
<td>Fair</td>
</tr>
<tr>
<td>.41 - .60</td>
<td>Moderate</td>
</tr>
<tr>
<td>.61 - .80</td>
<td>Substantial</td>
</tr>
<tr>
<td>.81 - 1.00</td>
<td>Almost perfect</td>
</tr>
</tbody>
</table>

Figure 7. *Sample size in observer agreement studies*

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Body Location</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karabulut et al.</td>
<td>1995</td>
<td>Oral</td>
<td>100</td>
</tr>
<tr>
<td>Abbey et al.</td>
<td>1995</td>
<td>Oral</td>
<td>120</td>
</tr>
<tr>
<td>Duray et al.</td>
<td>1992</td>
<td>Melanoma</td>
<td>50</td>
</tr>
<tr>
<td>Duncan et al.</td>
<td>1993</td>
<td>Melanoma</td>
<td>60</td>
</tr>
<tr>
<td>Piepkorn et al.</td>
<td>1994</td>
<td>Melanoma</td>
<td>60</td>
</tr>
<tr>
<td>Ismail et al.</td>
<td>1989</td>
<td>Cervical</td>
<td>100</td>
</tr>
<tr>
<td>Henrica et al.</td>
<td>1990</td>
<td>Cervical</td>
<td>106</td>
</tr>
<tr>
<td>Henrica et al.</td>
<td>1991</td>
<td>Cervical</td>
<td>93</td>
</tr>
<tr>
<td>Creagh et al.</td>
<td>1995</td>
<td>Cervical</td>
<td>122</td>
</tr>
<tr>
<td>Jones et al.</td>
<td>1996</td>
<td>Cervical Cytology</td>
<td>100</td>
</tr>
<tr>
<td>O'Sullivan et al.</td>
<td>1995</td>
<td>Cervical Cytology</td>
<td>110</td>
</tr>
<tr>
<td>Dixon et al.</td>
<td>1988</td>
<td>Colitis</td>
<td>100</td>
</tr>
<tr>
<td>Jensen et al.</td>
<td>1995</td>
<td>Colorectal</td>
<td>183</td>
</tr>
<tr>
<td>Carter et al.</td>
<td>1994</td>
<td>Anal</td>
<td>100</td>
</tr>
<tr>
<td>Fenger et al.</td>
<td>1990</td>
<td>Colorectal</td>
<td>56</td>
</tr>
<tr>
<td>Brown et al.</td>
<td>1985</td>
<td>Colorectal</td>
<td>100</td>
</tr>
</tbody>
</table>
Figure 8 - Study design flowsheet

1. Monday available cases (97 original slides) with 4 microinvasive cases & equal numbers of: normal, mild, moderate, severe/insitu

2. Each assessed by 3 pathologists
   - original referral description

3. 50% of cases reexamined by 3 pathologists (unchanged)

4. 50% of cases reexamined by 3 pathologists
   * altered referral description

5. Compare with MODE

6. Compare with SIGN-OUT

7. Interexaminer Agreement

8. Intraexaminer Agreement

9. Effect of Expectation of Prior Prevalence

10. Conformity

11. Consistency

12. Prior Prev

77
## Appendix 1. Summary of Dysplasia Observer Agreement Studies

<table>
<thead>
<tr>
<th>Author (Reference)</th>
<th>Year</th>
<th>N Cases</th>
<th>N Obs</th>
<th># Cats</th>
<th>Agreement</th>
<th>Inter</th>
<th>Intra</th>
<th>Pathologist Experience</th>
<th>Category Specific Agreement</th>
<th>Comparison Standard</th>
<th>Information Provided</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Oral</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Karabulut et al. (22)</td>
<td>1995</td>
<td>100</td>
<td>4</td>
<td>5</td>
<td>K=0.27-0.45</td>
<td>N/R</td>
<td>N/R</td>
<td>N/S between oral and general pathologists</td>
<td>N/R</td>
<td>N/R</td>
<td>Sex, age, topography</td>
</tr>
<tr>
<td>Abbey et al. (55)</td>
<td>1995</td>
<td>120</td>
<td>6</td>
<td>4</td>
<td>K= 0.15-0.40</td>
<td>K=0.05 - 0.49</td>
<td>All ≥10 years path experience</td>
<td>N/R</td>
<td>Signout diagnosis reviewed by 4 experts</td>
<td>None given</td>
<td></td>
</tr>
<tr>
<td>Lummerman et al. (27)</td>
<td>1995</td>
<td>744</td>
<td>2</td>
<td>5</td>
<td>%Agree= 54%</td>
<td>N/R</td>
<td>N/R</td>
<td>Both experienced</td>
<td>N/R</td>
<td>Full referral description</td>
<td></td>
</tr>
<tr>
<td>Lippman et al. (28)</td>
<td>1993</td>
<td>38</td>
<td>2</td>
<td>4</td>
<td>%Agree= 92%</td>
<td>N/R</td>
<td>N/R</td>
<td>Both experienced</td>
<td>N/R</td>
<td>Full referral description</td>
<td></td>
</tr>
<tr>
<td>Abbey et al. (46)</td>
<td>1998</td>
<td>120</td>
<td>6</td>
<td>4</td>
<td>AvgK=0.52</td>
<td>All ≥10 years path experience</td>
<td>N/R</td>
<td>Signout diagnosis reviewed by 4 experts</td>
<td>None given</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Nevi</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duray et al. (73)</td>
<td>1992</td>
<td>50</td>
<td>5</td>
<td>2</td>
<td>K=0.32-0.71</td>
<td>N/R</td>
<td>All experienced</td>
<td>N/R</td>
<td>Consensus by 3 dermatologists</td>
<td>None given</td>
<td></td>
</tr>
<tr>
<td>Piepkorn et al. (74)</td>
<td>1994</td>
<td>149</td>
<td>6</td>
<td>3</td>
<td>K= 0.22-0.54</td>
<td>K= 0.50-0.70</td>
<td>All experts</td>
<td>N/R</td>
<td>N/R</td>
<td>N/R</td>
<td></td>
</tr>
<tr>
<td>Duncan et al. (75)</td>
<td>1993</td>
<td>60</td>
<td>5</td>
<td>3</td>
<td>K= 0.05-0.047</td>
<td>N/R</td>
<td>3 experienced</td>
<td>K=0.38-0.47 Kpa=0.63-0.71 Kpa=0.55-0.84</td>
<td>N/R</td>
<td>N/R</td>
<td>N/R</td>
</tr>
<tr>
<td><strong>Colon</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pathologist B Intra K=0.9996</td>
<td>N/R</td>
<td></td>
</tr>
<tr>
<td>Jensen et al. (66)</td>
<td>1995</td>
<td>187</td>
<td>3</td>
<td>3</td>
<td>Kg= 0.33 Kg= 0.58</td>
<td>Kg= 0.42 Kg= 0.67</td>
<td>All experienced</td>
<td>N/R</td>
<td>N/R</td>
<td>N/R</td>
<td></td>
</tr>
<tr>
<td>Carter et al. (76)</td>
<td>1994</td>
<td>100</td>
<td>5</td>
<td>5</td>
<td>K= 0.09-0.48 K= 0.17-0.60</td>
<td>N/R</td>
<td>All experienced</td>
<td>N/R</td>
<td>N/R</td>
<td>N/R</td>
<td></td>
</tr>
<tr>
<td>Author (Reference)</td>
<td>Year</td>
<td>N Cases</td>
<td># Obs</td>
<td>Agreement</td>
<td>Pathologist Experience</td>
<td>Category Specific Agreement</td>
<td>Comparison Standard</td>
<td>Information Provided</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>------</td>
<td>---------</td>
<td>-------</td>
<td>-----------</td>
<td>------------------------</td>
<td>-----------------------------</td>
<td>----------------------</td>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cervix</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ismail et al. (77)</td>
<td>1989</td>
<td>100</td>
<td>7</td>
<td>7</td>
<td>Kg = 0.35 Kg = 0.55 Kgw = 0.72 Kg = 0.85</td>
<td>N/R</td>
<td>Category Inter/Intra K: Normal- K = 0.30/0.41 Reactive- K = 0.42/0.58 CIN1- K = 0.12/0.33 CIN2- K = 0.19/0.46 CIN3- K = 0.57/0.79 Invasive- K = 0.80/0.95</td>
<td>N/R</td>
<td>N/R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>De Vet et al. (5)</td>
<td>1990</td>
<td>106</td>
<td>4</td>
<td>5</td>
<td>Kg = 0.28 Kgw = 0.56</td>
<td>N/R</td>
<td>All experienced None= 35% Mild= 54% Moderate= 50% Severe= 49% Cis=22% All grades equally hard to distinguish.</td>
<td>N/R</td>
<td>N/R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>De Vet et al. (52) (after calibration)</td>
<td>1992</td>
<td>93</td>
<td>4</td>
<td>5</td>
<td>Kg = 0.32 Kgw = 0.69</td>
<td>N/R</td>
<td>All experienced</td>
<td>N/R</td>
<td>N/R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creagh et al. (53)</td>
<td>1995</td>
<td>122</td>
<td>5</td>
<td>6</td>
<td>Kg = 0.23 Kgw = 0.50</td>
<td>N/R</td>
<td>Normal K = 0.01 HPV K = 0.33 Borderline K = 0.06 Mild K = 0.20 Moderate K = 0.28 Severe K = 0.52</td>
<td>N/R</td>
<td>N/R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kato et al. (24)</td>
<td>1995</td>
<td>883</td>
<td>3</td>
<td>5</td>
<td>Kg = 0.56 Kgw = 0.58</td>
<td>N/R</td>
<td>All experienced</td>
<td>Normal K = 0.39 Mild K = 0.12 Moderate K = 0.16 Severe K = 0.58 Invasive K = 0.74</td>
<td>Panel consensus</td>
<td>Original description only provided to export panel if consensus could not be obtained</td>
<td></td>
</tr>
</tbody>
</table>
Bibliography


41) Sinton JM, Wood RE, Pharoah MJ, Lewis DW. Influence of the addition of restorations on


