Published in final edited form as:
Oral Surg Oral Med Oral Pathol Oral Radiol. 2015 October ; 120(4): 474–482.e2. doi:10.1016/j.oooo.2015.05.023.

Inter-Observer Agreement in Dysplasia Grading: Towards an Enhanced Gold Standard for Clinical Pathology Trials

Paul M. Speight, PhD, BDS, FDSRCPS, FDSRCS (Eng), FDSRCS (Edin.), FRCPath[a], **Timothy J. Abram, MS**[b], **Pierre N. Floriano, PhD**[d], **Robert James**[e], **Julie Vick, CCRP**[e], **Martin H. Thornhill, MBBS, BDS, PhD, MSc, FDSRCS (Edin), FDSRCSI, FDSRCS (Eng)**[f], **Craig Murdoch, BSc, PhD**[f], **Christine Freeman, BDS, FDSRCS (Eng), MMedSci, MPhil**[f], **Anne M Hegarty, BA, BDentSci, MSc, MBBS, MFD RCSI, FDS (OM) RCS**[g], **Katy D'Apice**[g], **A. Ross Kerr, DDS, MSD**[h], **Joan Phelan, DDS**[h], **Patricia Corby, DDS, MS**[i], **Ismael Khouly, DDS, MS, PhD**[i], **Nadarajah Vigneswaran, DMD, Dr. Med. Dent.**[j], **Jerry Bouquot, DDS, MSD**[j], **Nagi M. Demian, DDS, MD**[k], **Y. Etan Weinstock, MD**[l], **Spencer W. Redding, DDS, Med**[m], **Stephanie Rowan, RN, MSN**[m], **Chih-Ko Yeh, BDS, PhD**[m], **H. Stan McGuff, DDS**[n], **Frank R. Miller, MD, FACS**[o], and **John T. McDevitt, PhD**[b,c,p,**]

[a]Academic Unit of Oral & Maxillofacial Pathology, University of Sheffield School of Clinical Dentistry, Sheffield, UK

[b]Rice University, Department of Bioengineering, Houston, TX, USA

[c]Rice University, Department of Chemistry, Houston, TX, USA

[d]University of Texas MD Anderson Cancer Center, Houston, TX, USA

[e]Rho Inc., Chapel Hill, NC, USA

[f]Academic Unit of Oral & Maxillofacial Medicine & Surgery, University of Sheffield School of Clinical Dentistry, Sheffield, UK

[g]Unit of Oral Medicine, Charles Clifford Dental Hospital, Sheffield Teaching Hospitals National Health Service Foundation Trust, Sheffield, UK

[h]New York University College of Dentistry, Department of Oral and Maxillofacial Pathology, Radiology & Medicine, New York, NY, USA

[i]New York University College of Dentistry, Bluestone Center for Clinical Research, New York, NY, USA

[j]The University of Texas Health Science Center at Houston, Department of Diagnostic and Biomedical Sciences, Houston, TX, USA

[**]Corresponding Author: John T. McDevitt, Ph.D., Chair, Department Biomaterials, Bioengineering Institute, New York University, 433 First Avenue, Room 820, New York, NY 10010-4086, USA, mcdevitt@nyu.edu, Phone: 212-998-9204.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

DISCLOSURES

Principal Investigator, John T. McDevitt has an equity interest in SensoDX, LLC. and also serves on the Scientific Advisory Board.

[k]The University of Texas Health Science Center at Houston, Department of Oral and Maxillofacial Surgery, Houston, TX, USA

[l]The University of Texas Health Science Center at Houston, Department of Otolaryngology-Head and Neck Surgery, Houston, TX, USA

[m]The University of Texas Health Science Center at San Antonio, Department of Comprehensive Dentistry, San Antonio, TX, USA

[n]The University of Texas Health Science Center at San Antonio, Department of Pathology, San Antonio, TX, USA

[o]The University of Texas Health Science Center at San Antonio, Department of Otolaryngology-Head and Neck Surgery, San Antonio, TX, USA

[p]Department Biomaterials, Bioengineering Institute, New York University, 433 First Avenue, Room 820, New York, NY 10010-4086, USA

Abstract

Objective—Inter-observer agreement in the context of oral epithelial dysplasia (OED) grading has been notoriously unreliable and can impose barriers for developing new molecular markers and diagnostic technologies. This paper aimed to report the details of a 3-stage histopathology review and adjudication process with the goal of achieving a consensus histopathologic diagnosis of each biopsy.

Study Design—Two adjacent serial histological sections of oral lesions from 846 patients were independently scored by two different pathologists from a pool of four. In instances where the original two pathologists disagreed, a third, independent adjudicating pathologist conducted a review of both sections. If a majority agreement was not achieved, the third stage involved a face-to-face consensus review.

Results—Individual pathologist pair kappa values ranged from 0.251 – 0.706 (fair – good) before the 3-stage review process During the initial review phase, the two pathologists agreed on a diagnosis for 69.9% of the cases. After the adjudication review by a third pathologist, an additional 22.8% of cases were given a consensus diagnosis (agreement of 2 out of 3 pathologists). Following the face-to-face review, the remaining 7.3% of cases had a consensus diagnosis.

Conclusion—The use of the defined protocol resulted in a substantial increase (30%) in diagnostic agreement and has the potential to improve the level of agreement for establishing gold standards for studies based on histopathologic diagnosis.

INTRODUCTION

Cancers of the lip, oral cavity, and oropharynx are among the most common cancers, with approximately 400,000 incident cases globally.[1] In the United States, five year survival rates are approximately 60%, yet when diagnosed in the early stages, and confined to the primary site, the rates are 82%.[2] Oral squamous cell carcinomas (OSCC) comprise over 95% of cases and may be preceded by a precancerous lesion. The term oral potentially malignant disorder (OPMD) describes clinically detected epithelial lesions that carry an increased risk of progressing to cancer.[3] OPMDs range in clinical presentation from white patches (“leukoplakia”) to red patches (“erythroplakia”) or may be mixed red/white epithelial lesions (“erythroleukoplakia”) with or without ulceration. Following biopsy, these lesions can be graded on the basis of the histopathologic findings ranging from benign epithelial hyperkeratosis at one end of the spectrum, through mild, moderate, and severe oral epithelial dysplasia (OED), to carcinoma in situ, and finally OSCC at the other extreme.[4]

Histopathological assessment of OPMD depends on the microscopic grading of OED. However, such assessment can be subjective and notoriously unreliable with both poor inter- and intra-observer agreement between pathologists when assessing histological features.[5, 6] The challenges here may be traced to the fact that grading must impose artificial categories onto what is a diffuse, nonhomogeneous continuum of biological change, with no clear boundaries. Thus, it is challenging to define precise and reproducible criteria to categorize lesions to each side of these artificial boundaries. In general, grading systems work well when considering high grade dysplastic lesions or malignancy, but perform poorly for low grade dysplastic lesions. For these low grade lesions, changes may be subtle and there is considerable overlap with inflammatory and reactive histologic changes.[7]

Studies of grading systems used for other sites in the head and neck region (e.g., larynx) have shown similar weaknesses[8], often due to lack of consensus on the best grading systems as well as to inter-observer variability. One common method of quantifying inter-observer agreement is by calculating kappa values. The interpretation of kappa statistics throughout this paper, including previously published kappa values, is based upon the scale proposed by Altman (1991): <=0.20 = poor, 0.21 – 0.40 = fair, 0.41 – 0.60 = moderate, 0.61 – 0.80 = good, 0.81 – 1.00 = very good.[9] In their review, Fleskens et al. (2009)[8] noted that published kappa values for inter-observer agreement of oral dysplasia varied from 0.17 (poor) to 0.78 (good). They also found that, similar to other sites, grading of oral dysplasia was more valid and reproducible for severe dysplasia and malignancy than it was for low grade lesions. A number of studies have evaluated inter- and intra-observer agreement in the grading of dysplasia and all have shown variable degrees of agreement which are better for high grade lesions and for pathologists trained in the same institution.[10–13] Abbey et al. (1995) studied the degree of agreement between 6 pathologists grading 120 lesions. This study revealed that the average inter-observer agreement with the original diagnosis of the presence or absence of dysplasia was only 81.8% with kappa values ranging from 0.29 to 0.57 (fair to moderate). The intra-observer agreement by pathologists with their own previous grading of dysplasia averaged 81.4% (kappa values ranging from 0.31 to 0.71). Put another way, when using established histologic criteria, pathologists were not able to confirm their own previous opinion that dysplasia was present or absent in nearly 20% of cases.[10]

In an attempt to resolve these issues, Kujan et al. (2006) evaluated a binary grading scheme which used the World Health Organization (WHO) morphological criteria to categorize dysplastic lesions into either ‘low-risk’ or ‘high-risk’.[14] Kujan and co-workers determined that ‘high-risk’ lesions, which subsequently underwent malignant transformation, were characterized by at least four architectural changes and at least five cytological changes. ‘Low-risk’ lesions, which did not progress, showed less than four architectural changes or less than five cytological changes. The authors then used these criteria to grade 68 lesions using either the recommended 5-point scheme (WHO[4]) or their proposed binary scheme. Both schemes were found to predict malignant transformation with good correlations between 'low-risk' and hyperplasia and mild dysplasia, and between 'high-risk' and severe dysplasia or carcinoma in situ. However, the binary scheme was better able to categorize lesions showing moderate dysplasia into high- and low-risk groups, with 14 of 16 designated as high risk progressing to cancer. The binary scheme also showed good discrimination for predicting progression-free survival. However, the overall kappa values for inter-observer agreement were similar in both schemes and were only fair to moderate (WHO grading system: 0.22, Binary grading system: 0.50).

In spite of these shortcomings, grading of dysplasia remains the most valid method for assessing the malignant potential of OPMD.[15]

With the goal of creating new quantitative and un-biased tools with the potential to aid in the diagnosis and management of patients with OPMDs, we have assembled a team of clinicians who are involved in the diagnosis and management of OPMDs, alongside a group of experienced bioengineers and cancer biologists. This multi-disciplinary team has just completed the recruitment phase of a prospective 999-patient phase 2/3 clinical trial to validate oral cancer biomarker signatures derived from quantitative cytological and immunohistochemical image-based parameters, facilitated by a novel lab-on-a-chip sample processing approach. A major first step in the study was to agree on the histopathologic criteria for the diagnosis and grading of OED, in order to determine a method for the establishment of enhanced gold standard diagnoses against which the cytological and biomarker expression profiles would be compared.

This paper reports the details of the multistage histopathology review and adjudication process. This program followed a well-defined protocol that was designed with the objective of improving the level of agreement among pathologists in the microscopic assessment of OPMD. The use of this defined protocol has the potential to improve the level of agreement for studies where the histopathologic assessment sets the gold standard diagnosis.

MATERIALS AND METHODS

The study reported here is part of a larger study designed to evaluate a new chip-based system (lab-on-a-chip) to measure cytological parameters on brush biopsy samples to assist in the diagnosis and management of OPMD. The study was a prospective non-interventional trial involving a single visit by patients who presented with OPMD. The study was conducted at four sites: i) the University of Texas Health Science Center at Houston, ii) the University of Texas Health Science Center at San Antonio, iii) Bluestone Center for Clinical Research at New York University, and iv) Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield in the United Kingdom. The study was approved by the Institutional Review Boards of all participating institutions, including that of Rice University where chip-based measurements were completed on brush biopsy samples. The data from the chip-based measurements will be reported in future publications and are outside the scope of this paper which will report the results of the protocol to establish agreement on the histopathologic diagnosis of the lesions.

All patients provided written informed consent. Rho Inc., a contract research organization (Chapel Hill, North Carolina), provided statistical, regulatory, data management, and clinical monitoring support, as well as operational management. Throughout the trial, all data that was collected was entered in a web-based Electronic Data Capture (EDC) system.

PATIENT GROUPS

Patients attending Oral Medicine, Oral Surgery, or Otolaryngology clinics at the participating institutions were recruited into the study. A total of 999 subjects were enrolled into three study groups:

Group 1—The major group consisted of 775 patients who had OPMDs and who underwent a scalpel biopsy as part of the normal standard of care for diagnosis of the lesion.

Group 2—The second group comprised 74 patients with OSCC that had been diagnosed by a prior scalpel biopsy and confirmed histopathologically within 45 days of enrollment.

Group 3—The third group was a total of 150 lesion-free, healthy volunteers. Participants in this group provided brush biopsy samples of their tongue and buccal mucosa; healthy volunteers did not undergo a scalpel biopsy.

CLINICAL PROTOCOL

Brush Biopsy—Patients in Group 1 underwent brush biopsy of the oral lesion (OPMD) and also a brush biopsy of the contralateral, clinically normal mucosa. The brush biopsy sample was taken immediately before the same lesion underwent a scalpel biopsy as described below. Patients in Group 2 underwent brush biopsy of the known cancerous lesion, as well as the contralateral, clinically normal mucosa. For healthy volunteers in Group 3, a brush biopsy of normal appearing tissue on the lateral or ventral surface of the tongue and a brush biopsy of normal appearing tissue on the left or right buccal mucosa were taken. Brush biopsy samples were taken using a soft Rovers® Orcellex® oral cytology brush (Rovers Medical Devices B.V., Oss, The Netherlands). The brush was applied directly to the lesion or control oral mucosa using mild pressure and rotated 360° approximately 10–15 times in the same direction to obtain the cytological sample.

Scalpel Biopsy—As part of clinical patient management, a scalpel biopsy was performed on Group 1 subjects with oral mucosal lesions suspicious for OPMD following standard clinical procedures appropriate for each individual case. Patients in Group 2, who previously underwent a diagnostic biopsy, were not subjected to an additional scalpel biopsy.

LABORATORY PROTOCOL

Scalpel Biopsy Processing and Histopathological Diagnosis—Tissue from scalpel biopsies on Group 1 subjects were formalin fixed, paraffin embedded and processed for routine histopathologic examination. For patient management purposes, histopathologic diagnosis was made by the attending pathologists of the respective institutions following their standard procedures.

Brush Biopsy Processing—At the clinical site's laboratory, the brush biopsy samples were immediately processed, frozen, and stored in a −80°C freezer until they were sent in batches to the McDevitt Laboratory at Rice University according to standard operating protocols. The detailed protocols for preparation of the brush biopsy samples and the results of the biomarker analyses will be reported in subsequent papers.

Histopathological Analysis—For research study purposes, hematoxylin and eosin stained sections of the scalpel biopsy specimens were examined by the four participating study pathologists. These participating pathologists (PMS, JP, NV, SMcG) were all senior and experienced oral and maxillofacial pathologists. The qualifications of each pathologist can be found along with their institutional affiliations in the author list. In addition to the diagnostic slide used for patient management, two consecutive serial sections were prepared for each specimen and sent for research review by two pathologists (designated as reviewers A and B) blinded to the clinical and microscopic diagnosis, and to the site of the lesion. Reviewers A and B therefore reviewed different slides from the same lesion – but these were adjacent serial sections and therefore only approximately 5um apart. The four individual participating pathologists are denoted as i, ii, iii, and iv. The pairs of reviewing pathologists were designated "Reviewer A" and "Reviewer B" and were paired from the pool of the four participating pathologists. Reviewers A and B were at different centers from each other and also from the participating pathologist who issued the original diagnosis. Each pathologist independently categorized each case into one of 7 microscopic diagnostic categories based on the 2005 WHO guidelines for typing of cancer and precancer of the oral mucosa. The exact terminology and their respective microscopic descriptions used were agreed upon in advance by the pathologists participating in the trial, facilitated by the contract research organization that provided oversight for the study. Table 1 summarizes the diagnostic classifications used for the study. There was no attempt to calibrate the pathologists beyond the agreed terminology summarized in Table 1. For subjects with a previously diagnosed malignant lesion, slides generated from the initial biopsy that led to the diagnosis of malignancy were used to confirm the diagnosis.

Since the microscopic classification of the biopsies were to be used as a gold standard for the lab-on-chip machine learning and diagnostic model selection, research diagnoses of the biopsies needed to be as objective as possible. These were based solely on an analysis of the architectural and cytological changes following the WHO guidelines.[4] As such, when the research pathologists evaluated and classified the lesions into one of the 7 research diagnostic categories, they were blinded to the patient's clinical findings, the clinical impressions of those who had taken the biopsy, and the original histopathology report generated for patient management purposes.

ADJUDICATION PROCESS

To enable the research pathologists to reach agreement on the histopathological diagnosis of each biopsy, a 3-stage review process was implemented as shown in Figure 1. The process was completed as follows:

Adjudication—After the first independent examination of the sections, if pathologists A and B agreed, then this diagnosis was accepted as the gold standard diagnosis for those cases. Since the goal of the study design was to increase the level of inter-observer agreement, a pragmatic decision was made to accept an agreed upon diagnosis as final, which is the norm for most studies. If there was disagreement between the two pathologists, BOTH slides were reviewed by a third independent pathologist for adjudication. The same pathologist acted as the adjudicator for all cases. This pathologist (JB) was an experienced senior oral and maxillofacial pathologist who has published widely on the topic of oral epithelial dysplasia and premalignancy. He was independent of the previous review stages and was blinded to the clinical details, to the original diagnosis and to the opinions of reviewers A and B. A majority diagnosis (i.e., agreement by 2 out of the 3 pathologists (reviewer A, reviewer B and adjudicating pathologist)) was accepted as the enhanced gold standard diagnosis.

Consensus Review—For those cases where the adjudicator did NOT agree with either reviewer A or B, slides were subjected to a consensus review at which Reviewers A and B and the adjudicating pathologist met for a face-to-face meeting at UTHSC Houston. BOTH slides for each case were reviewed and discussed, and the three pathologists reached agreement on a diagnosis. Rather than isolating the three pathologists, the slides were reviewed and discussed as a group in order to maximize the probability of achieving an accurate final diagnosis. As with the initial reviews, diagnoses were based on histology alone. Clinical information was not provided to the pathologists. Additionally, the consensus review was performed independent of the initial histopathological assessment (i.e., reviewers were blinded to the original diagnoses and comments). Slides were reviewed using a multi-head microscope.

The diagnosis was recorded on a source document worksheet for each case undergoing consensus review. Using the provided worksheet, the new diagnosis was recorded separately from the initial diagnosis to ensure all original data was maintained. Following completion of the review, data were entered into the Electronic Data Capture system.

Binary Review—Using the previously published grading criteria, all cases were reclassified according to the method described by Kujan et al.[14] as a high-risk or low-risk lesion. High-risk lesions included all lesions graded as severe dysplasia or carcinoma in situ, and low-risk lesions included lesions graded as non-dysplastic or showing mild dysplasia. The binary classification review was undertaken to re-classify all cases graded as moderate dysplasia into the high- or low-risk categories.[14] The review was undertaken at the same time as the consensus review during a face-to-face meeting at UTHSC Houston. All cases with a final research diagnosis of moderate dysplasia were shipped to the UTHSC Houston site prior to the meeting and any cases designated as moderate dysplasia during the consensus review were subsequently included in the binary classification review.

Each moderate dysplasia case (two slides per subject) was reviewed by two pathologists. The two pathologists discussed and scored the approved architectural and cytological criteria[4, 14] and determined if the lesion should be categorized as high-risk or low-risk. Lesions with 3 or less architectural criteria or 4 or less cytological criteria were re-classified as low-risk. Lesions with 4 or more architectural criteria and 5 or more cytological criteria were re-classified as high-risk. As with the initial, adjudication, and consensus reviews, diagnoses were based on histology alone, and clinical information was not provided to the pathologists. The diagnosis was recorded on a source document worksheet for each case undergoing review. Using the provided worksheet, the new binary diagnosis was recorded separately from the initial, adjudication, and consensus diagnoses to ensure all original data were maintained. Because the study was designed to use the final diagnoses in assigning low-risk or high-risk status, the binary review process does not assess inter-observer agreement levels.

MATHEMATIC BASIS FOR ADJUDICATION PROCESS

Since the true diagnosis is unknown, it is not possible to determine the true rates of correct or incorrect diagnosis. However, by using basic probability theory, we have demonstrated the statistical benefit of the 3-stage adjudication process. Probabilities for correct and incorrect diagnoses were derived from the overall levels of agreement and disagreement between reviewers in the initial review stage. Details for the derivation can be found in the Supplementary Materials. The key assumptions made in this example were that: 1) all reviewers had an equal probability of misdiagnosis that was not influenced by the other reviewers, 2) each slide had an equal probability of misdiagnosis, and 3) where two reviewers disagree on a particular diagnosis, one was assumed correct and the other was assumed incorrect. Though there is a definite possibility that two disagreeing reviewers may both be incorrect regarding a particular diagnosis, this assumption is necessary to calculate the estimated probabilities.

The estimated probabilities for each scenario are shown in Table 2. The addition of an adjudicator in instances when two reviewers disagree on a diagnosis resulted in an increase in the total probability of a correct diagnosis from 66.4% to 91.0% (Appendix I). Even though simplifying assumptions were used to derive these probabilities, these calculations suggest the substantial reduction in diagnostic disagreement that can be gained by using our adjudication protocol.

RESULTS

Of the 849 recruited patients in groups 1 (OPMD lesions) and 2 (patients with previously diagnosed OSCC), a total of three patients were excluded from the pathologist agreement analysis due to one of their two slides being inadequate for diagnostic purposes, resulting in a total of 846 specimens for review. The initial review stage consisted of different pairs of pathologists. Separate agreement levels for these pairs of pathologists (Table 3) were calculated using the kappa statistic and percent agreement. Kappa statistics ranged from 0.251 to 0.706, translated as “fair agreement” to “good agreement” based on the interpretation set by Altman (1991).[9] The breakdown of observed reviewer agreement across the 7 diagnostic categories is shown in Table 4.

The two reviewers were in full agreement for 591 (69.9%) of the 846 cases that were reviewed across 7 diagnostic categories. When full agreement was not achieved, slides were reviewed by an independent pathologist as part of the adjudication stage. Following this stage, an additional 22.8% of cases were given a consensus diagnosis. For the remaining 7.3% of cases in which all three pathologists (the two reviewers and adjudicating pathologist) lacked agreement, a face-to-face consensus review was performed. Following consensus review, agreement was reached on the diagnosis for 100% of the 846 eligible cases.

Final classifications were determined throughout the 3-stage adjudication process whenever full agreement or majority agreement was reached. Of the 846 lesion samples, 545 (64.4%) were classified as benign lesions, 107 (12.6%) as mild epithelial dysplasia, 33 (3.9%) as moderate dysplasia, 16 (1.9%) as severe dysplasia, 3 (0.4%) as carcinoma in situ, and 142 as OSCC (16.8%). All moderate dysplasia diagnoses were further reviewed using a binary review, “low-risk/high-risk” classification system. Using this system, 662 (78.2%) of the 846 lesion samples were classified as “low-risk” and 184 (21.7%) were classified as “high-risk”. Of the 33 lesions classified as moderate dysplasia, 10 were re-classified as “low-risk” and 23 were re-classified as “high risk”.

DISCUSSION

Inter-observer disagreement has been a notorious consequence of dysplasia grading, resulting from the challenges of imposing artificial categories onto continuous biological changes. Greater levels of disagreement are found when categorizing low grade lesions.[7] This observation is likely due to the fact that these lesions show fewer and more subtle changes, many of which may be seen in reactive lesions, especially as a result of an inflammatory infiltrate. In our study, the majority of the dysplastic lesions were mild or low risk, which may account for the relatively low kappa values and low levels of inter-observer agreement after the initial review. This lack of agreement is particularly challenging in the context of defining robust, objective standards for training predictive diagnostic models. The use of our well-defined protocol resulted in an increase in pathologist agreement from 69.9% to 100%. The more confident histopathological diagnoses may serve as an enhanced gold standard for studies that validate early-stage diagnostic tools for OPMDs.

Besides one reviewer set (i/ii) exhibiting a high level of agreement with a kappa value of 0.706 and a percent agreement of 81%, the remaining 4 reviewer sets, with kappa values ranging from 0.251 – 0.513 and percent agreement ranging from 62%–79.6%, correlate well with previous studies evaluating inter-observer variability in the diagnosis and grading of OPMD where kappa values ranged from 0.15 – 0.70 with a percent agreement range of 35.8% – 69%.[10–14] It is worth noting that the sample size used in these studies ranged from 64 – 120 slides, while the sample size in this study ranged from 105 – 234 cases per pair of reviewers with a total of 846 cases. By applying the 3-stage adjudication process protocol, a final “enhanced gold standard” diagnosis was established for each slide. The foundation for the use of this adjudication protocol stems from the well-known phenomenon of the performance of collective judgments of groups compared to single expert opinions. This effect has been demonstrated across many diverse fields, including medical expert opinion.[19, 20] In his book, “The Wisdom of Crowds”, James Surowiecki states that groups can make better decisions than its individual members if the members are allowed to function independently.[21] By increasing the inter-observer agreement level, we hypothesize that the collective opinion of several pathologists may result in a more clinically accurate microscopic diagnosis. This *enhanced gold standard* may serve as a benchmark for future model training and for validation of novel molecular and morphometric biomarkers used in OED risk stratification.

DOES A GOLD STANDARD EXIST?

In a recent study that evaluated inter-observer agreement in histopathological grading of OED, Dost et al.[22] concluded that OED grading has such poor predictive value that it should not be used as a treatment guide. A follow-up editorial regarding this conclusion by Edwards (2014)[23] countered that, despite its limitations, OED grading gives a pathologist the best opportunity to convey the overall risk of malignancy to the clinician. Additionally, the authors agreed that molecular markers are needed to assist the pathologist and may eventually lead to a more definitive OED risk stratification.

An important distinction in terminology is made by Bosman (2001)[5] who classifies “research histopathology” and “applied histopathology” as two separate entities. Cross-platform comparisons can lead to poor levels of inter-observer agreement and clinical predictive value because the purposes of each category are different. We acknowledge that the derivation of an enhanced gold standard presented here should be considered only for research histopathology because its primary aim is to provide an objective benchmark for developing biomarker technologies to assist clinical diagnosis. The final translation of these new technologies into clinical practice would then be considered clinical histopathology, or in Bosman’s terminology ‘applied histopathology’.

For a scoring system to be clinically useful it must demonstrate high levels of both inter-observer and intra-observer reproducibility.[11, 24] Furthermore, inter- and intra-observer agreement levels can provide an estimate of the validity of grading systems when an appropriate gold standard is not available.[11] Due to the low levels of inter- and intra-observer agreement regarding histopathological OED grading presented in this study and throughout various literature reviews, there is a significant need for a reliable methodology that derives the highest level of agreement from the current, imperfect gold standard.

ENHANCED GOLD STANDARD FOR NEW TECHNOLOGY DEVELOPMENT

When developing a predictive model, such as a diagnostic tool based on molecular biomarker data, collected data is partitioned into two datasets: a training set and a testing set. A gold standard is required to generate known outcomes for both data sets in order to evaluate the performance of the new predictive model. While known outcomes for the training set are used to calibrate the model, users are blinded to the outcomes for the testing set in order to simulate a real world scenario. Therefore, the evaluation of the performance of a new predictive model is entirely dependent on the reliability of the gold standard used. Though the clinical accuracy cannot be determined without performing a patient outcome study of the presented adjudication process, we believe that by increasing the level of agreement between different observers, researchers can create a benchmark of comparison for their developing technologies.

The adjudication and consensus review process reported here were designed to increase inter-observer agreement, in order to achieve higher confidence in the reported histopathological assessments. An increase in agreement does not necessarily reflect an increase in diagnostic accuracy. However, evidence from our probabilistic case-study (Table 2) and the "wisdom of crowds" phenomenon support our hypothesis that the collective opinion of several independent pathologists has a greater probability of achieving more clinically accurate diagnoses. Correlation with clinical outcome still remains the most relevant measure of any risk stratification procedure[5, 25], but the intention of the adjudication process outlined here is to provide an intermediary benchmark for rapid, iterative development of novel biomarkers and their accompanying model training and validation. We have demonstrated that a substantial increase (30%) in diagnostic agreement can be accomplished by using the described adjudication and consensus process.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Rho, Inc. (Chapel Hill, North Carolina) for generously providing assistance with patient data management and statistical analysis. We would especially like to thank Vanessa Hearnden, BSc, PhD[f] for technical and administrative support and Deborah Holt, MBBS, BDS, PhD[f] for performing biopsies.

We would also like to the University of Texas Health Science Center at San Antonio (UTHSCSA), University of Texas Health Science Center at Houston (UTHSCH), New York University/Bluestone Center for Clinical Research, Sheffield Teaching Hospitals NHS Foundation Trust and the University of Sheffield for assistance in obtaining clinical samples.

Funding for this work was provided by the National Institutes of Health (NIH) through the National Institute of Dental and Craniofacial Research (Award Number 1RC2DE02078501, 5RC2DE020785-02, 3RC2DE020785-02S1, 3RC2DE020785-02S2). This content is solely the responsibility of the authors and does not necessarily represent or reflect the official views of the NIH or the US government. Segments of this work are supported by Renaissance Health Service Corporation and Delta Dental of Michigan.

References

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer. Dec 15; 127(12):2893–917. [PubMed: 21351269]
2. SEER Cancer Statistics Factsheets: Oral Cavity and Pharynx Cancer. Available from: http://seer.cancer.gov/statfacts/html/oralcav.html
3. Warnakulasuriya S, Johnson NW, van der Waal I. Nomenclature and classification of potentially malignant disorders of the oral mucosa. J Oral Pathol Med. 2007 Nov; 36(10):575–80. [PubMed: 17944749]
4. Barnes, L.; Eveson, JW.; Reichart, P.; Sidransky, D. World Health Organization classification of tumors: pathology and genetics of head and neck tumors. Lyon: IARC Press; 2005.
5. Bosman FT. Dysplasia classification: Pathology in disgrace? Journal of Pathology. 2001; 194(2): 143–4. [PubMed: 11400140]
6. Warnakulasuriya S, Reibel J, Bouquot J, Dabelsteen E. Oral epithelial dysplasia classification systems: predictive value, utility, weaknesses and scope for improvement. J Oral Pathol Med. 2008 Mar; 37(3):127–33. [PubMed: 18251935]
7. Montgomery E. Is there a way for pathologists to decrease interobserver variability in the diagnosis of dysplasia? Arch Pathol Lab Med. 2005 Feb; 129(2):174–6. [PubMed: 15679414]

8. Fleskens S, Slootweg P. Grading systems in head and neck dysplasia: their prognostic value, weaknesses and utility. Head & neck oncology. 2009; 1:11. [PubMed: 19432960]

9. Altman, D. Practical statistics for medical research. London: Chapman and Hall; 1991.

10. Abbey LM, Kaugars GE, Gunsolley JC, Burns JC, Page DG, Svirsky JA, et al. Intraexaminer and interexaminer reliability in the diagnosis of oral epithelial dysplasia. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 1995 Aug; 80(2):188–91. [PubMed: 7552884]

11. Brothwell DJ, Lewis DW, Bradley G, Leong I, Jordan RCK, Mock D, et al. Observer agreement in the grading of oral epithelial dysplasia. Community Dentistry and Oral Epidemiology. 2003; 31(4):300–5. [PubMed: 12846853]

12. Fischer DJ, Epstein JB, Morton TH, Schwartz SM. Interobserver reliability in the histopathologic diagnosis of oral pre-malignant and malignant lesions. J Oral Pathol Med. 2004 Feb; 33(2):65–70. [PubMed: 14720191]

13. Karabulut A, Reibel J, Therkildsen MH, Praetorius F, Nielsen HW, Dabelsteen E. Observer variability in the histologic assessment of oral premalignant lesions. J Oral Pathol Med. 1995 May; 24(5):198–200. [PubMed: 7616457]

14. Kujan O, Oliver RJ, Khattab A, Roberts SA, Thakker N, Sloan P. Evaluation of a new binary system of grading oral epithelial dysplasia for prediction of malignant transformation. Oral Oncology. 2006; 42(10):987–93. [PubMed: 16731030]

15. Napier SS, Speight PM. Natural history of potentially malignant oral lesions and conditions: an overview of the literature. J Oral Pathol Med. 2008 Jan; 37(1):1–10. [PubMed: 18154571]

16. Holmstrup P. Can we prevent malignancy by treating premalignant lesions? Oral Oncology. 2009; 45(7):549–50. [PubMed: 18952490]

17. Holmstrup P, Vedtofte P, Reibel J, Stoltze K. Oral premalignant lesions: is a biopsy reliable? J Oral Pathol Med. 2007 May; 36(5):262–6. [PubMed: 17448135]

18. Kelloff GJ, Boone CW, Crowell JA, Nayfield SG, Hawk E, Malone WF, et al. Risk biomarkers and current strategies for cancer chemoprevention. Journal of Cellular Biochemistry. 1996; 63(SUPPL 25):1–14. [PubMed: 8891900]

19. Robson N, Rew D. Collective wisdom and decision making in surgical oncology. European Journal of Surgical Oncology (EJSO). 36(3):230–6. [PubMed: 20106625]

20. Gillis CR, Hole DJ. Survival outcome of care by specialist surgeons in breast cancer: A study of 3786 patients in the west of Scotland. British Medical Journal. 1996; 312(7024):145–8. [PubMed: 8563532]

21. Surowiecki, J. The Wisdom of Crowds. New York: Anchor Books; 2004.

22. Dost F, LÃª Cao K, Ford PJ, Ades C, Farah CS. Malignant transformation of oral epithelial dysplasia: A real-world evaluation of histopathologic grading. Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology. 2014; 117(3):343–52.

23. Edwards PC. The natural history of oral epithelial dysplasia: Perspective on Dost et al. Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology. 2014; 117(3):263–6.

24. Cross SS. Grading and scoring in histopathology. Histopathology. 1998; 33(2):99–106. [PubMed: 9762541]

25. Lessells AM, Burnett RA, Goodlad JR, Howatson SR, Lang S, Lee FD, et al. Comment on a recent paper and editorial on the subject of dysplasia classification. Journal of Pathology. 2002; 198(1): 131–2. [PubMed: 12210073]

STATEMENT OF CLINICAL RELEVANCE

In an effort to increase inter-observer agreement amongst pathologists in oral epithelial dysplasia grading, we have developed a 3-stage review and adjudication protocol with the goal of enabling an enhanced gold standard for new technology development.

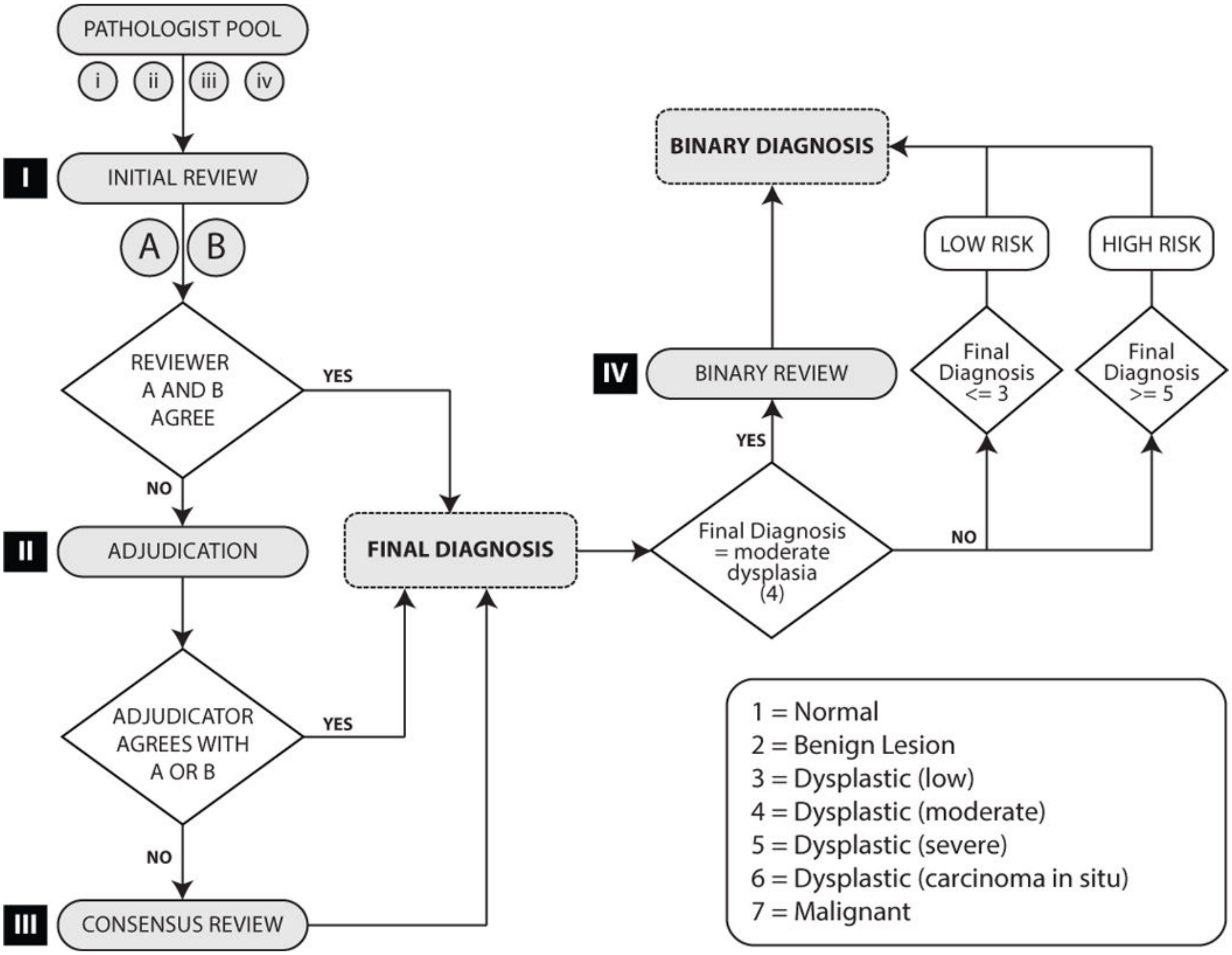

Figure 1.

Flow chart illustrating the process for the enhanced gold standard adjudication sequence.

Table 1

Definition of diagnostic outcomes, including normal epithelium and the full spectrum of potentially malignant disorders.

Diagnosis	**Histopathologic Criteria**
Non-neoplastic/normal	Surface stratified squamous epithelium demonstrates normal thickness without hyperplasia or hyperkeratinization. The underlying lamina propria is devoid of chronic inflammatory cell infiltrate.
Benign lesions	Surface stratified squamous epithelium may reveal hyperkeratosis and hyperplasia, but without cellular atypia and disordered maturation (dysplasia). The underlying lamina propria may exhibit chronic inflammatory cell infiltrate: Chronic mucositis. This category may encompass a range of benign lesions including benign hyperkeratosis and lichen planus.
Dysplastic (mild)	Surface stratified squamous epithelium reveals cellular atypia and disordered maturation (dysplasia) limited to the basal and parabasal layers or verruciform epithelial hyperplasia and hyperkeratosis with mild degree of atypical architecture.
Dysplastic (moderate)	Surface stratified squamous epithelium reveals cellular atypia and disordered maturation (dysplasia) extending from the basal layer to the mid portion of the spinous layer or verruciform epithelial hyperplasia and hyperkeratosis with moderate degree of atypical architecture.
Dysplastic (severe)	Surface stratified squamous epithelium reveals cellular atypia and disordered maturation (dysplasia) extending from the basal layer to a level above the midpoint of the epithelium or verruciform epithelial hyperplasia and hyperkeratosis with severe degree of atypical architecture.
Dysplastic (carcinoma *in situ*)	Surface stratified squamous epithelium reveals cellular atypia and disordered maturation (dysplasia) involving the entire thickness of the epithelium.
Malignant	Islands and cords of malignant squamous epithelial cells arise from dysplastic surface epithelium and invade into the lamina propria.

Table 2

Probability of correct 7-level diagnosis (normal, benign, dysplastic mild, dysplastic moderate, dysplastic severe, dysplastic carcinoma in situ, malignant) with two reviewers and the use of an adjudicator when the two reviewers disagree. The values of Pc and Pw represent the probability of a correct and wrong diagnosis, respectively. Even though the model assumptions area a simplification of reality, these probabilities so derived do suggest the substantial gains in correctly diagnosing the lesions that are likely to be achieved through our adjudication process.

Probability Scenario	**Reviewer A**	**Reviewer B**	**Adjudicator**	**Probability**		**Final Diagnosis**
				Equation	**Probability**	
1	Correct	Correct	N/A	Pc *Pc	0.664	Correct
2	Wrong	Wrong	N/A	Pw * Pw	0.034	Wrong
3	Correct	Wrong	Correct	Pc * Pw * Pc	0.123	Correct
4	Correct	Wrong	Wrong	Pc* Pw * Pw	0.028	Wrong
5	Wrong	Correct	Correct	Pw * Pc * Pc	0.123	Correct
6	Wrong	Correct	Wrong	Pw * Pc * Pw	0.028	Wrong
Total Probability of Correct Diagnosis					0.91 (91.0%)	
Total Probability of Wrong Diagnosis					0.090 (9.0%)	
Overall Total Probability					1.00 (100%)	

Table 3

Agreement reviewing pathologists during Initial Review Stage. Initial Review Stage percent agreement and kappa values are shown for individual pathologist pairs.

Initial Review				
Reviewing Pathologists	**N**	**Kappa (Interpretation)**	**Kappa 95% CI**	**% Agreement**
i | ii	147	0.706 (Good)	(0.618, 0.793)	81.0%
i | iii	245	0.513 (Moderate)	(0.427, 0.600)	79.6%
ii | iii	115	0.251 (Fair)	(0.126, 0.377)	65.2%
iii | iv	105	0.463 (Moderate)	(0.336, 0.589)	68.6%
ii | iv	234	0.423 (Moderate)	(0.339, 0.508)	62.0%

Table 4

Level of agreement in lesion diagnoses after Initial Review Stage. Cells contain the count of slide pairs with the column label showing the diagnosis of one reviewer and the row label showing the diagnosis of the other reviewer. The actual identity of the reviewer is not reflected in this table; rather it reports all of the observed diagnostic combinations after the Initial Review Stage. The diagonal cells (gray) contain the count of diagnostic pairs where both reviewers were in complete agreement in their diagnosis. For example: the cell where the “Benign” column crosses with the “Dysplastic Mild” row represents 128 subject biopsies where one reviewer diagnosed one slide as benign and the other reviewer diagnosed the other slide as dysplastic mild.

Percentage Agreement 69.9%	Normal	Benign	Dysplastic Mild	Dysplastic Moderate	Dysplastic Severe	Dysplastic Carcinoma in situ	Malignant
Normal	0						
Benign	1	408					
Dysplastic Mild	0	128	31				
Dysplastic Moderate	0	41	38	14			
Dysplastic Sever	0	8	4	16	5		
Dysplastic Carcinoma in situ	0	0	0	1	1	3	
Malignant	1	5	3	2	5	1	130