Risk Stratification System for Oral Cancer Screening

Lutécia H. Mateus Pereira[1]*, Isildinha M. Reis[2, 3]*, Erika P. Reategui[4], Claudia Gordon[2], Sandra Saint-Victor[4], Robert Duncan[3], Carmen Gomez[5], Stephanie Bayers[6], Penelope Fisher[4], Aymee Perez[2], W. Jarrard Goodwin[2, 4], Jennifer J. Hu[2, 3], Elizabeth J. Franzmann[2, 4]

[1]Miami-Dade College, Miami, FL, USA

[2]Sylvester Comprehensive Cancer Center, University of Miami Leonard Miller School of Medicine, Miami, FL, USA

[3]Department of Public Health Sciences, University of Miami Leonard Miller School of Medicine, Miami, FL, USA

[4]Department of Otolaryngology, University of Miami Leonard Miller School of Medicine, Miami, FL, USA

[5]Department of Pathology, University of Miami Leonard Miller School of Medicine, Miami, FL, USA

[6]MacNeal Hospital, Berwyn, IL, USA

* Participated equally as first authors

Running Title: Salivary Biomarkers for Oral Cancer

Key Words: squamous cell carcinoma, oral cavity, oropharynx, CD44, protein, early detection

Financial Support: E.J. Franzmann received: NCI R01CA118584, NCI RO3 CA107828, 4BB-20 Bankhead-Coley, 10BG-02 Bankhead-Coley. The work was also funded by Woman's Cancer Association, a gift from Vigilant Biosciences, Inc., Sylvester

Comprehensive Cancer Center and University of Miami, Department of Otolaryngology.

Corresponding Author: Elizabeth J. Franzmann, Department of Otolaryngology, University of Miami Leonard Miller School of Medicine, Clinical Research Building Room 1513, 1120 NW 14th St, Miami, FL 33136 Phone: 305-243-5955

email: efranzman@med.miami.edu

Disclaimer: The University of Miami, Drs. Franzmann, Reis, Pereira and Duncan hold intellectual property used in the study and have potential for financial benefit from its future commercialization. Dr. Franzmann is Chief Scientific Officer, consultant, and equity holder in Vigilant Biosciences, Inc., licensee of the intellectual property used in this study.

Word Count: 3,924

Total Number of Tables: 3 main document, 4 supplementary

Total Number of Figures: 2

ABSTRACT

Oral cavity and oropharyngeal cancer (oral cancer) is a deadly disease that is increasing in incidence. World-wide 5-year survival is only 50% due to delayed intervention with more than half of diagnoses at stage III and IV, whereas earlier detection (stage I and II) yields survival rates up to 80%-90%. Salivary soluble CD44 (CD44), a tumor-initiating marker, and total protein levels may facilitate oral cancer risk assessment and early intervention. This study used a hospital-based design with 150 cases and 150 frequency-matched controls to determine whether CD44 and total protein levels in oral rinses were associated with oral cancer independent of age, gender, race, ethnicity, tobacco and alcohol use, and socioeconomic status (SES). High-risk subjects receiving oral cancer prevention interventions as part of a community- based program (n=150) were followed over 1 year to determine marker specificity and variation. CD44 ≥5.33 ng/ml was highly associated with case status (adjusted OR 14.489, 95%CI: 5.973, 35.145; p<.0001, versus reference group CD44 <2.22 ng/ml and protein <1.23 mg/ml). Total protein aided prediction above CD44 alone. Sensitivity and specificity in the frequency-matched study was 80% and 48.7%, respectively. However, controls were not representative of the target screening population due, in part, to a high rate of prior cancer. In contrast, specificity in the high-risk community was 74% and reached 95% after annual retesting. Simple and inexpensive salivary CD44 and total protein measurements may

help identify individuals at heightened risk for oral cancer from the millions who partake in risky behaviors.

INTRODUCTION

Head and neck squamous cell carcinoma (HNSCC), which includes cancers of the oral cavity, pharynx and larynx, affects 550,000 people world-wide each year (1). In India, oral cancer, defined here as cancers of the oral cavity and oropharynx, is the most common fatal cancer in middle-aged men, and it is the costliest cancer in low-income countries (2,3). The main risk factors include tobacco use, alcohol use, and human papillomavirus (HPV) infection (4-6). The incidence of oral cancer is rising with the increasing incidence of HPV+ oropharyngeal cancer (7).

World-wide 5-year survival only reaches 50%, largely due to late stage (III or IV) presentation (8). Upper aerodigestive tract (UADT) mucosa progresses through a premalignant phase, dysplasia, prior to development of frank malignancy. Dysplasia is reversible (9) and can regress with tobacco cessation or spontaneously (10,11). Unfortunately, dysplasia often mimics characteristics of benign inflammation so, frequently, it remains occult until further progression results in late stage cancer diagnosis (12).

Screening for HNSCC in India reduced oral cancer mortality by over 80% in tobacco and/or alcohol users (13). Screening by oral exam followed by tissue biopsy, the gold standard, has only 64% sensitivity for oral cancer (8) and 31% specificity for oral dysplasia or cancer (14).

Molecular tests including hypermethylation, RNA, and protein-based panels are under development, but not validated (15-18). Other technologies that use dyes, autofluorescence or exfoliative cytology as adjuncts to the physical exam are used in clinical practice but have not improved early detection rates (19,20).

CD44, a cell surface transmembrane glycoprotein involved in cell proliferation, cell migration, and tumor initiation (21-24) is overexpressed in premalignant lesions (25-27). Soluble CD44 (solCD44), released by proteinases, is detectable in body fluids (28,29). Prior work suggests that total protein is also an effective oral cancer marker (30,31). Both can be measured with simple, inexpensive assays and are overexpressed in oral cavity and oropharyngeal cancers suggesting usefulness in both HPV positive and negative disease (29-32).

This study uses a case-control, hospital-based design to evaluate salivary markers in oral cancer cases and controls, frequency-matched for important risk and demographic factors to determine whether CD44 and total protein levels are associated with cancer rather than potential confounders. The markers are then tested in a community at elevated risk for oral cancer (n=150) at baseline and 1-year follow-up to examine marker changes over time. Moreover, this study begins to explore whether oral rinse CD44 and total protein levels 1) detect both HPV positive and HPV negative disease, 2) are associated with prognosis, and 3) change over a 1 year period. The outcome of this work is

a reliable, inexpensive and noninvasive risk prediction test for oral cancer with potential to greatly benefit populations that suffer most from this disease.

MATERIALS and METHODS

Case control design to determine marker cutpoints

Subjects for the 2012 hospital-based, case-control study were recruited from clinics at the University of Miami Sylvester Comprehensive Cancer Center (UM) and Jackson Memorial Hospital (JMH) between 2007 and 2012 (Figure 1). This study evaluated whether soluble salivary tumor markers distinguish 150 oral cancer patients from 150 controls frequency matched for age, gender, race, ethnicity, tobacco and alcohol use, and socioeconomic status (SES). Oral cancer cases included newly diagnosed, previously untreated subjects with squamous cell carcinoma. Control subjects were identified from family medicine and internal medicine clinics and chosen, prior to testing so that the key covariates (age, tobacco use etc.) in the control group were not significantly different from the covariates in the case group. All subjects were recruited equally from UM, a private university hospital system serving mostly insured, white patients and JMH, a county hospital system serving primarily low--income patients and a large minority population. All subjects completed a questionnaire including demographics, behavioral risk factors and SES. For cases, data on tumor characteristics and outcomes were extracted from medical records. Controls with lesions suspicious for oral cancer were excluded as were HIV

positive or pregnant individuals. Exclusion decisions were blinded to marker level results. The resulting marker panel was validated using 27 oral cavity and oropharyngeal cases and 39 high-risk controls enrolled between 2004 and 2006 in a previous case-control study (31).

Test performance in a high-risk target screening group

The hospital-based, case-control study was designed to determine whether CD44 and total protein were associated with oral cancer independent of demographic and risk variables. To determine the specificity of the markers in a potential target screening population, the marker panel developed using data from the 2012 hospital-based, case-control study was evaluated in 150 participants from a community previously determined to be at elevated risk for oral cancer due to poverty and smoking (33). Subjects in this study received free head and neck cancer screening, education on smoking cessation, good nutrition and oral health. This community control group was followed over time; baseline and annual follow-up oral rinses were obtained and measured between the years 2011 and 2013 to assess specificity and variation in marker levels. Since the community control group was still at elevated risk for cancer, we also estimated true specificity in a group of 21 normal volunteers who were primarily nonsmokers.

All participants consented according to The Code of Ethics of the World Medical Association (Declaration of Helsinki).

Laboratory Analysis

Oral rinses were collected using a previously published method that samples the oral cavity and oropharynx (29-32). Levels of solCD44 (normal and variant isoforms) were measured using a sandwich ELISA assay (eBioscience), with previously published modifications (29-32). We performed the DC protein assay (Bio-Rad Laboratories) according to the manufacturer's protocol using saliva samples prepared as previously published (29-32). Each sample was tested in duplicate and the technician was blinded to disease status.

Formalin-fixed and paraffin-embedded specimens were retrieved from cases, where available (n=79). HPV status was assessed using $p16^{INK4A}$ immunohistochemistry (IHC), an accepted surrogate marker for HPV (34-36). $p16^{INK4A}$ was performed according to the manufacturer's IHC protocol on 68 specimens (BD Bioscience). Additionally, HPV status was already available in 11 cases (IHC, n=10 or in situ hybridization, n=1). All specimens were reviewed by a pathologist (CG), blinded to the patient's clinical data. $p16^{INK4A}$ expression was scored as positive if strong and diffuse nuclear and cytoplasm staining was present in ≥ 50% of the tumor specimen (36).

Statistical Analysis

Patient groups were compared with respect to the distribution of potentially important categorical covariates using the chi-square test or Fisher's exact test. Data on solCD44 were log base-2 transformed to stabilize estimates

of variance and improve the fit to the normal distribution. Continuous variables were analyzed using Student's t-test or analysis of variance (ANOVA) followed by Fisher's least-significant-difference test for pairwise mean comparison, and tests of pre-specified contrasts. Logistic regression analysis was used to assess the association between markers and the risk for oral cancer. Odds ratio (OR) estimates were reported with corresponding 95% confidence interval (95%CI) and area under the curve (AUC) of the receiver operating characteristic curve (ROC) for fitted models. Estimates of sensitivity, specificity, and accuracy were derived from a fitted multivariable logistic model which included significant interactions between markers and covariates as well as from a model including only risk groups based on cutpoints for solCD44 and protein that were derived using multivariate recursive partitioning analysis (37) implemented in the R-packages MVPART (v.1.6.1.) and Recursive Partitioning and Regression Trees (RPART), version 1.6-0 (38). Kaplan-Meier and Cox regression models were used to evaluate progression-free survival (PFS) and overall survival (OS). Hazard ratio (HR) estimates and corresponding 95%CI are reported. Statistical analyses were performed using SAS version 9.2 (SAS Institute, Inc.) and R package.

RESULTS

Characteristics of Hospital--Based Case-Control Study

The description of the hospital-based, case-control study, comprising 150 patients with oral cancer and 150 controls, is summarized in Table 1 and Figure 1. There were no significant differences between cases and controls with respect to age, gender, race, SES, oral health (number of teeth removed), smoking history, alcohol habit or enrollment clinic (county JMH versus private hospital UM system). Supplemental Table 1 (online version only) shows cancer specific characteristics for cases.

Log_2solCD44, hereafter referred to as CD44, and total protein were evaluated with respect to risk factors or demographic variables within the case and control groups (Table 2). CD44 and protein were higher in cases compared to controls at the $p<0.05$ level when age, gender, race/ethnicity, SES, smoking habit or drinking habit, teeth loss or ability to gargle were considered. This provides strong evidence that CD44 and total protein levels are associated with oral cancer independent of these risk factors. In cases but not in controls, CD44 was significantly higher in subjects who were older, had worse gargle ability, and more teeth loss. CD44 and protein did not differ significantly by TNM status or HPV status.

HPV+ tumors, frequent in nonsmokers with oropharyngeal HNSCC, have a better prognosis compared to smoking and alcohol related tumors (39). HPV+ tumors are rarely found in the OC. In our study, only 4 out of the 31 HPV + tumors were from the OC (see supplemental Table 1, online only). The CD44 levels between the 4 OC HPV+ and 27 HPV+ OP cases were not significantly

different. The total protein levels were significantly lower (OC: 0.54 mg/ml OP: 0.93 mg/ml p=0.001) in the OC compared to OP HPV+ samples.

Risk Modeling

In univariate analysis, CD44 and total protein were significantly associated with cancer status with an OR for 1-unit increase in CD44 of 2.036 (95%CI: 1.552, 2.671, p<0.0001, AUC=0.68) and for 1-unit increase in protein of 2.159 (95%CI: 1.288, 3.617, p<0.003, AUC=0.59). The AUC was improved to 0.763 in a multivariable model including adjustments for important variables and their interactions, which removed residual confounding not accounted for in the frequency matching. The OR for CD44 increased to 2.684 (95%CI: 1.797, 4.010, p<.0001), while the OR for protein became less than 1 and non-significant (OR=0.646, 95%CI: 0.301, 1.386, p=0.262) (Table 3, part A). This model "markers + covariates" with AUC=0.763 provided significantly better prediction than the reduced model excluding both markers and only including potential risk factors (AUC=0.686) (p=0.003), indicating that the markers aid prediction over and above prediction provided by knowledge of risk factors alone.

Findings for the analysis stratified by $p16^{INK4A}$ (surrogate for HPV status) were similar to the combined analysis. In the HPV- group, protein levels were associated with a significant protective effect following multivariate analysis (Table 3, part B).

Multivariate recursive partitioning and logistic regression analyses were employed to understand the relationship between CD44, protein and prediction of disease presence (Table 3, part C). Importantly, when covariates including CD44, protein, age, gender, race, ethnicity, and SES, were included into the model, CD44 and protein were the most important predictors of cancer status, defining 5 risk groups. Furthermore, we found that the AUC=0.722 for the risk group model defined by CD44 and protein is significantly different from AUC=0.681 for the univariate log2 CD44 model (p=0.025), indicating that the addition of protein improves prediction.

The classification tree defined subjects as “controls” if CD44 was <2.22 ng/ml and protein was <1.23 mg/ml (reference group) or if CD44 was ≥2.22 & <5.33 ng/ml and protein was ≥0.558 mg/ml (Table 3, part C). However, compared to reference group, the odds ratio for the latter group was 2.192 (95%CI:1.247, 3.854) and significant (p=0.006), indicating elevated risk. Furthermore, many cancer subjects and 2 control subjects who went on to develop cancer during the course of the study had levels that fell into this medium CD44 and medium-high protein group leading us to consider this group as a case group. The other groups classified as “cases” included subjects with CD44 <2.22 ng/ml and protein ≥1.23 mg/ml, CD44 ≥2.22 & <5.33 ng/ml and protein <0.558 mg/ml, and CD44 ≥5.33 ng/ml, regardless of protein level. Thus, based on the levels of CD44 and total protein, we identified 4 of the 5 groups at risk as cases. Odds ratios derived from a multivariate model including risk

groups defined by CD44 and protein, demographic and risk factors showed similar results (Table 3, part C). The percentage of cancer patients that fell into each risk category by HPV status and stage is shown in Online Version Only -Supplementary Table 2.

Defining the reference group as control and all others as cases, sensitivity was 80.7% and specificity was 48.7% for the 2012 hospital-based group (see Table 3, Part C for numbers of cases and controls that fell into each group). Sensitivity reached 80% for Stages I-IV, and 85% for Stage I-II. These results were validated using CD44 and total protein results from a similar hospital-based study whose enrollment was completed in 2006 (single test, stage I-IV: sensitivity- 2012=80.7%, 2006=77.8%; specificity- 2012=48.7%, 2006=56.4%). The frequency-matched control group was at exceptionally high risk for cancer since over 10% of these controls had a history of prior cancer outside the UADT. Hospital-based controls with history of cancer had significantly higher solCD44 and protein levels compared to controls without prior cancer ($p<0.05$). Thus, the community-based population was used to estimate the specificity of the test. This is in keeping with suggestions by the Early Detection Research Network who note that control subjects from clinical settings may not be representative of control subjects recruited from the population because they have been referred for some reason to the clinic (40). They suggest that, although selection based on convenience may be necessary

early, final conclusions should be based on population--based studies, if possible (40).

Specificity in a Target Screening Population

To predict specificity in a true screening population, a community at high-risk for HNSCC (n=150, see supplemental Table 3 (online version only) for demographic and risk characteristics) were evaluated. These subjects were all African-American, they were heavier smokers and drinkers and had worse oral health than the cases. They were younger than the cases and were enrolled from a community center rather than a clinic. We also studied oral rinses from 21 normal volunteers. Specificity was greatest in the normal volunteers (95.2%). Specificity was 74% (n=150) after one baseline evaluation but also reached 95% in the high--risk community in subjects retested at one year (n=95). In the latter case, a result was considered positive if both the baseline and annual result were positive. Importantly, these subjects had received counseling on smoking cessation, nutrition and oral health and assistance with access to such services as part of the oral cancer prevention program prior to this apparent drop in marker levels.

Changes in CD44 and protein levels over time in a screening population

A total of 95 patients in the community-based control group provided baseline and annual follow-up collections. The distribution of changes in CD44 and protein over 1 year is shown in Figure 2A and B, respectively. The average annual drop in CD44 of -0.439 ng/ml (24%) was significant ($p<.0001$). Linear

regression analysis confirmed a significant linear trend for lower CD44 values (R^2=0.227, intercept=0.785 (p<.0001), slope=0.331 (p<.0001)), (Figure 2C). Mean protein also dropped from 0.644 to 0.543 mg/ml (p=0.036) with confirmation by linear regression analysis (R^2=0.108, intercept=0.284 (p=0.002), slope=0.402 (p<.0001), (Figure 2D). Of 22 community subjects at baseline elevated risk only 5 remained in an at-risk category after 1 year follow-up suggesting that retesting may improve specificity.

To determine if the decreased marker levels were due to variation in assay conditions over the course of the year rather than a true decrease in the markers, a baseline second aliquot (baseline 2) was run on the same plate as the annual follow-up collection with 81 such pairs for each assay (protein and CD44). The average drop in levels between baseline 2 and annual follow-up was significant for CD44 (CD44: -0.296ng/ml, p=0.023; protein: -0.013 mg/ml, p=0.796) while linear regression showed a significant trend towards lower numbers for both markers (CD44: R^2=0.227, intercept=0.882 (p<.0001), slope=0.288 (p<.0001); protein. R^2=0.155, intercept=0.256 (p=0.008), slope=0.534 (p<.0001); figures not shown). We also fit linear regression of baseline 2 on baseline 1. For CD44, linear regression indicated that the two baselines were equivalent suggesting that the changes in CD44 level were not due to technical changes in the assay. The differences between baselines for protein were not within the expected random variation (data not shown).

Prognostic Significance of Markers

Overdiagnosis has been observed in breast, prostate and thyroid cancer screening (41). To avoid this, markers should identify aggressive forms of oral cancer rather than indolent cancers that will not cause significant problems during a patient's lifetime (41). We assessed marker association with prognostic factors and adjusted for confounders such as stage to determine whether the markers have potential to detect early, aggressive forms of the disease. Kaplan-Meier curves for progression-free survival (PFS) and overall survival (OS) by risk group are shown in Figure 2, E and F. Unadjusted and adjusted estimates of hazard ratios (HRs) for PFS and OS by risk groups are shown in Supplemental Table 4 (online version only). Based on multivariate analysis with adjustment for tumor stage, age, gender, race and ethnicity, and SES, hospital-based cases that had CD44 levels ≥5.33 ng/ml, had reduced PFS (adjusted HR=3.919, 95%CI: 1.692, 9.080, p=0.001) and OS (adjusted HR=3.242, 95%CI: 1.299, 8.089, p=0.012) compared with cases in the reference group. Subjects with CD44 <2.22 ng/ml and protein ≥1.23 mg/ml had borderline association with decreased PFS (adjusted HR=3.446, 95%CI: 0.857, 13.867, p=0.082) and no significant difference in OS (adjusted HR=2.186, 95%CI: 0.524,9.123, p=0.284) compared to cases in the reference group; however, this group included only 4 cases. As a result, the data supports the markers indeed have potential to identify the most aggressive forms of oral cancer.

Potential application of CD44 in detecting oral cancer or cancers at other sites

Similar to prior work (32), in this study, 2 control subjects fell into an elevated risk category and developed early HNSCC (lip and carcinoma in situ of the larynx) in follow-up. Two other controls were excluded because of bladder cancer and possible oral pre-malignancy, respectively. The latter, went on to develop lung cancer. A subject from the community-based study classified in an elevated risk category developed lung cancer 14 months following testing.

DISCUSSION

Despite over 550,000 new diagnoses of HNSCC worldwide each year, few receive a skilled oral cancer screening exam. Early diagnosis dramatically improves survival, but most present late. This study describes a simple, inexpensive, noninvasive risk assessment test based on salivary CD44 and protein that is able to distinguish stage I-IV oral cancer cases from controls. Sensitivity of early stage lesions was as good or better (I-II=85%) than identification of all stages combined (I-IV=80%). The finding that early and late stage disease is detected is in keeping with prior publications on salivary CD44 levels by this and other groups (32,42).

Also consistent with prior work, adding total protein increases the accuracy of the test at very minimal cost (30). The relative protein and CD44 levels may greatly facilitate risk stratification since these specific levels are associated with varying risk, as indicated by the odds ratio. This may enable clinicians to tailor follow-up and patients to understand their risk better thus

motivating change. Further work must be done to determine the cutpoints that characterize multiple risk levels across diverse populations.

A strength of the study that adds to prior work is frequency-matching which ensures that there are no statistically significant differences between cases and controls with respect to age, race, gender, SES, tobacco use or alcohol use. This ensures that the biomarkers are associated with cancer risk and not some other confounder such as tobacco use. We included additional covariates in modeling to remove any residual confounding. Results strongly support that CD44 and total protein are associated with cancer risk independent of tobacco or alcohol use, age gender, race, etc.

In frequency-matching, a hospital-based control group was chosen since the cases were also hospital-based and the goal was to ensure that the cases and controls were as similar as possible except for cancer status. While this minimizes confounding, there are some limitations to this design as control subjects from such clinics may not be representative of control subjects recruited from the population (40). Indeed, over 10% of our control population had a prior history of cancer. The EDRN suggests that final conclusions should be based on population--based studies, if possible (40). To begin to investigate the markers in a population-based control group, we also enrolled a target screening group of smokers from an underserved, minority community population and followed them over time. When we compared results to the

cases from the hospital-based study, specificity in the high-risk community could reach as high as 95% after annual retesting, increasing from 74% for a single initial test. Given that this population had higher levels of tobacco and alcohol use, worse oral health and lower SES than the case group, this specificity is quite high.

The study included a diverse population. We enrolled subjects from a public institution that serves primarily unfunded patients and includes a large percentage of Hispanic and African-American minorities as well as a private academic institution that serves mostly funded, white patients. Thus we had ample minorities, patients from low SES, and patients with poor oral health. This further ensures that the markers will work in diverse populations and limits potential confounding.

The study provides exploratory evidence that high salivary CD44 is associated with poor PFS and OS. Thus, CD44 appears to be associated with aggressive disease, though further study would be needed to determine whether these markers are useful for prognosis (43).

We performed preliminary analysis on CD44 and protein levels in HPV+ versus HPV- cancer. Our data did not show a significant difference in CD44 or protein levels between HPV+ and HPV- subjects. We do not think it is related to oral cavity cases that were HPV+ since there were only 4 of these and the CD44 levels were not significantly different than the oropharyngeal HPV+ samples. Protein levels were significantly higher in the OP HPV+ compared to OC HPV+

subjects, though the sample size was small. HPV status was unknown in 39 of 91 of the oropharyngeal cases. This is a limitation of the study thus further investigation is needed better understand the relationship between CD44, total protein and HPV status.

Two false positive control subjects developed HNSCC during the follow-up period. Additional subjects with other smoking-associated tumors, including lung and bladder, also had elevated CD44 levels. Thus, “false positives” could actually be true positives for occult oral disease or other cancers. Since CD44 is a tumor initiation factor, levels might go down if risk factors decrease and occult lesions disappear. Data suggests that individuals who stayed in the community screening program for a year underwent a significant decrease in CD44 levels not attributable to technical differences in the test. All subjects who stayed in the community screening program received education on smoking cessation and access to resources to assist them in improving oral hygiene and nutrition raising the possibility that these prevention efforts may result in lower marker levels and lower risk. However, more investigation is needed to show this definitively.

This study assesses risk of oral cancer in that certain levels of CD44 and protein are associated with elevated ORs and the OR for relatively rare diseases like oral cancer approximates relative risk (44). While the study does provide directional, anecdotal evidence that certain levels of CD44 and protein may identify those patients that will go on to develop cancer or precancer, the study

was not designed to assess leukoplakia or dysplasia or determine whether these markers predict progression to invasive cancer. Whether these markers predict progression is an area of considerable interest that should be explored further in larger, prospective studies.

Conclusion: The results provided here are encouraging. Further investigations with larger sample sizes are needed to determine whether marker levels vary with behavioral changes such as smoking cessation, whether reversal of premalignant lesions is associated with a drop in marker levels, and whether this test increases the number of screen detected oral cancer lesions. Success in any of these areas could revolutionize oral cancer screening, by providing a simple and reliable measure of oral cancer risk that alerts primary care providers, dentists and other frontline screeners to individuals most in need of skilled oral exam at a stage when the process can be more easily treated or perhaps even reversed with behavioral modification.

Acknowledgements: We would like to thank members of the University of Miami, Division of Head and Neck Surgery, the Department of Family Medicine, the Sylvester Comprehensive Cancer, Center Disparities and Community Outreach Core, Liberty Square Community Center, Curley's House Food Bank, Liberty City Community Health Advisory Board and Dr. John Deo for their assistance with this work.

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 2010;127:2893-917.

2. Dikshit R, Gupta PC, Ramasundarahettige C, Gajalakshmi V, Aleksandrowicz L, Badwe R, et al. Cancer mortality in India: a nationally representative survey Lancet 2012; 379: 1807–16.

3. The global economic cost of cancer. American Cancer Society, Inc. 2010.

4. Gillison ML, Broutian T, Pickard RK, Tong ZY, Xiao W, Kahle L, et al. Prevalence of oral HPV infection in the United States, 2009-2010. JAMA 2012;307:693-703.

5. D'Souza G, Kreimer AR, Viscidi R, Pawlita M, Fakhry C, Koch WM, et al. Case-control study of human papillomavirus and oropharyngeal cancer. N Engl J Med 2007;356:1944-56.

6. Blot WJ, McLaughlin JK, Winn DM, Austin DF, Greenberg RS, Preston-Martin S, et al. Smoking and drinking in relation to oral and pharyngeal cancer. Cancer Res 1988;48:3282-7.

7. American Cancer Society. Cancer Facts and Figures 2012. Atlanta: American Cancer Society 2012.

8. Brocklehurst P, Kujan O, Glenny AM, Oliver R, Sloan P, Ogden G, et al. Screening programmes for the early detection and prevention of oral cancer. Cochrane database of systematic reviews 2010:CD004150.

9. Pindborg JJ, Daftary DK, Mehta FS. A follow-up study of sixty-one oral dysplastic precancerous lesions in Indian villagers. Oral Surg Oral Med Oral Pathol 1977;43:383-90.

10. Larsson A, Axéll T, Andersson GJ. Reversibility of snuff dippers' lesion in Swedish moist snuff users: a clinical and histologic follow-up study. Oral Pathol Med. 1991;20:258-64.

11. Grizzle WE, Srivastava S, Manne U. The biology of incipient, pre-invasive or intraepithelial neoplasia. Cancer Biomark 2010;9:21-39.

12. Poh CF, Ng SP, Williams PM, Zhang L, Laronde DM, Lane P, et al. Direct fluorescence visualization of clinically occult high-risk oral premalignant disease using a simple hand-held device. Head Neck 2007;29:71-6.

13. Sankaranarayanan R, Ramadas K, Thara S, Muwonge R, Thomas G, Anju G, et al. Long-term effect of visual screening on oral cancer incidence

and mortality in a randomized trial in Kerala, India. Oral Oncol 2013;49:314-21.

14. Epstein JB, Güneri P, Boyacioglu H, Abt E. The limitations of the clinical oral examination in detecting dysplastic oral lesions and oral squamous cell carcinoma. J Am Dent Assoc 2012;143:1332-42.

15. Hu S, Arellano M, Boontheung P, Wang J, Zhou H, Jiang J, et al. Salivary proteomics for oral cancer biomarker discovery, Clin Cancer Res 2008;14;6246-52

16. Carvalho AL, Jeronimo C, Kim MM, Henrique R, Zhang Z, Hoque MO, et al. Evaluation of promoter hypermethylation detection in body fluids as a screening/diagnosis tool for head and neck squamous cell carcinoma. Clin Cancer Res 2008;14:97-107.

17. Elashoff D, Zhou H, Reiss J, Wang J, Xiao H, Henson B, et al. Prevalidation of salivary biomarkers for oral cancer detection. Cancer Epidemiol Biomarkers Prev 2012;21:664-72.

18. Cheng YL, Rees T, Wright J. A review of research on salivary biomarkers for oral cancer detection. Clin Transl Med 2014;3:3.

19. Lingen MW, Kalmar JR, Karrison T, Speight PM. Critical evaluation of diagnostic aids for the detection of oral cancer. Oral Oncol 2008;44:10-22.

20. Balevi B. Assessing the usefulness of three adjunctive diagnostic devices for oral cancer screening: a probabilistic approach. Community Dent Oral Epidemiol 2011;39:171-6.

21. Screaton GR, Bell MV, Jackson DG, Cornelis FB, Gerth U, Bell JI. Genomic structure of DNA encoding the lymphocyte homing receptor CD44 reveals at least 12 alternatively spliced exons. Proc Natl Acad Sci USA 1992;89:12160-4.

22. Ponta H, Sherman L, Herrlich PA. CD44: from adhesion molecules to signaling regulators. Nature Rev Mol Cell Biol 2003;4:33-45.

23. Perez A, Neskey DM, Wen J, Pereira L, Reategui EP, Goodwin WJ, et al. CD44 interacts with EGFR and promotes head and neck squamous cell carcinoma initiation and progression. Oral Oncol 2013;59:306-313

24. Prince ME, Sivanandan R, Kacsorowski A, Wolf GT, Kaplan MJ, Dalerba P, et al. Identification of a subpopulation of cells with cancer stem cell properties in head and neck squamous cell carcinoma. Proc Natl Acad Sci USA 2007;104:973-8.

25. Hirvikoski P, Tammi R, Kumpulainen E, Virtaniemi J, Parkkinen JJ, Tammi M, et al. Irregular expression of hyaluronan and its CD44 receptor is associated with metastatic phenotype in laryngeal squamous cell carcinoma. Virchows Arch 1999;434:37-44.

26. Ioachim E, Assimakopoulos D, Goussia AC, Peschos D, Skevas A, Agnantis NJ. Glycoprotein CD44 expression in benign, premalignant and malignant epithelial lesions of the larynx: an immunohistochemical study including correlation with Rb, p53, Ki-67 and PCNA. Histol Histopathol 1999;14:1113-8.

27. Dasari S, Rajendra W, Valluru L. Evaluation of soluble CD44 protein marker to distinguish the premalignant and malignant carcinoma cases in cervical cancer patients. Med Oncol 2014;31:139.

28. Kajita M, Itoh Y, Chiba T, Mori H, Okada A, Kinoh H, et al. Membrane-type 1 matrix metalloproteinase cleaves CD44 and promotes cell migration. J. Cell Biol 2001;153:893-904.

29. Franzmann EJ, Reategui EP, Carraway KL, Hamilton KL, Weed DT, Goodwin WJ. Salivary soluble CD44: a potential molecular marker for head and neck cancer. Cancer Epidemiol Biomarkers Prev 2005;14:735-9.

30. Franzmann EJ, Reategui EP, Pereira LH, Pedroso F, Joseph D, Allen GO, et al. Salivary protein and solCD44 levels as a potential screening tool for early detection of head and neck squamous cell carcinoma. Head Neck 2012;34:687-95.

31. Pereira LH, Adebisi IN, Perez A, Wiebel M, Reis I, Duncan R, et al. Salivary markers and risk factor data: a multivariate modeling approach for head and neck squamous cell carcinoma detection. Cancer Biomark 2011;10:241-9.

32. Franzmann EJ, Reategui EP, Pedroso F, Pernas FG, Karakullukcu BM, Carraway KL, et al. Soluble CD44 is a potential marker for the early detection of head and neck cancer. Cancer Epidemiol Biomarkers Prev 2007;16:1348-55.

33. Dietz NA, Sherman R, Mackinnon J, Fleming L, Arheart KL, Wohler B, et al. Toward the identification of communities with increased tobacco-associated cancer burden: Application of spatial modeling techniques. J Carcinog 2011; 10:22

34. Klussmann JP, Gultekin E, Weissenborn SJ, Wieland U, Dries V, Dienes HP, et al. Expression of p16 protein identifies a distinct entity of tonsillar carcinomas associated with human papillomavirus. Am J Pathol 2003;162:747-53.

35. Hafkamp HC, Manni JJ, Haesevoets A, Voogd AC, Schepers M, Bot FJ, et al. Marked differences in survival rate between smokers and nonsmokers with HPV 16-associated tonsillar carcinomas. Int J Cancer 2008;122:2656-64.

36. El-Naggar AK, Westra WH. p16 expression as a surrogate marker for HPV-related oropharyngeal carcinoma: a guide for interpretative relevance and consistency. Head Neck 2012;34:459-61.
37. Breiman, L, Friedman, JH, Olshen, RA, Stone, CJ. (1984). Classification and Regression Trees. Wadsworth, Belmont, CA.
38. Therneau TM, Atkinson B. 2014. "MVpart" A package for running multivariate regression trees in R software. URL http://cran.r-project.org/web/packages/mvpart/.
39. Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tân PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med 2010;363:24-35.
40. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001;93:1054-61.
41. Srivastava S, Reid BJ, Ghosh S. Kramer BS. Research needs for understanding the biology of overdiagnosis in cancer screening. J of Cell Physiol 2015 Oct 27 [Epub ahead of print]
42. Allegra E, Trapasso S, Sacco A, Aragona T, Belfiore A, and Garozzo A. Elisa Detection of Salivary Levels of Cd44sol as a Diagnostic Test for Laryngeal Carcinomas. Otolaryngology-Head and Neck Surgery. J Cancer Sci Ther 2012; 4:330-4.

43. Chen J, Zhou J, Lu J, Xiong H, Shi X, Gong L . Significance of CD44 expression in head and neck cancer: a systemic review and meta-analysis. BMC Cancer 2014;14:15.

44. Bonita R., Beaglehole R, Kjellström T. (2006) Basic Epidemiology 2nd Ed. Geneva: World Health Organization.

Table 1. Characteristics of cases and controls

Variable / Category	Cases (n=150)		Controls (n=150)		*P*
	N	%	N	%	
Site of enrollment					
JMH	80	53.3	71	47.3	0.299
UM	70	46.7	79	52.7	
Age, years					
<40	4	2.7	-	-	0.214
40 |- <50	20	13.3	29	19.3	
50 |- <60	60	40.0	56	37.3	
60 |- <70	44	29.3	44	29.3	
≥70	22	14.4	21	14.0	
<60	84	66.0	85	56.7	0.449
≥60	66	44.0	65	43.3	
Mean (Standard deviation)	58.6 (10.5)		58.5 (9.7)		0.887
Median (Range)	58 (28 - 88)		58.5 (40 – 87)		
Gender					
Male	121	80.7	118	78.7	0.907
Female	29	19.3	32	21.3	
Race					

White	123	82.6	118	79.7	0.534
Black	26	17.4	30	20.3	
Asian/Other/Missing (1 case Other, 1 control Asian, and 1 control missing)	1		2		
Ethnicity					
Hispanic	77	51.3	93	62.0	0.062
Non-Hispanic	73	48.7	57	38.0	
SES[1]					
Low	100	66.7	90	60.0	0.231
High	50	33.3	60	40.0	
Oral health score					
Poor/Fair	80	64.0	87	58.0	0.310
Good	45	36.0	63	42.0	
Missing	25				
Teeth removed					
None/1 to 5	86	58.9	92	63.0	0.301
6 or more but not all	36	24.7	39	26.7	
All	24	16.4	15	10.3	
Missing	4		4		
Smoking status					
Never	33	22.0	32	21.3	0.889
Ever	117	78.0	118	78.7	

Drinking habits[2]					
Non-drinker/Mild	78	52.0	85	56.7	0.279
Moderate	24	16.0	30	20.0	
Heavy	48	32.0	35	23.3	

[1] Socioeconomic status (SES) categories high and low were defined based on income (≤$25,000, >$25,000), education ("≤grade 12 or GED", "some college or college graduate") and employment ("out-of/unable-to work", "occupation with some income"). (See online Version Only -Supplementary Table 3)**. High SES**: income >$25,000 or, if income was missing, "some college or college graduate" AND "occupation with some income". **Low SES**: income ≤$25,000, or, if income was missing, low education and/or "out-of/unable-to work"; 1 subject missing income and education with "occupation with some income" was classified as low SES.

[2] **Drinking habits: Non-drinker/Mild**: past drinking ≤2 drinks/day or current drinking ≤2 drinks/day for 1-15 days/month; **Moderate**: past drinking 3 to <5 drinks/day or current drinking ≤2 drinks/day for 16-30 days/month or ≥3 drinks/day for 1-15 days/month; **Heavy**: past drinking 5 or more drinks/day or current drinking ≥3 drinks /day for 16-30 days/month.

Table 2. log_2solCD44, and protein levels in oral rinses of R01 HNSCC study by patient group and key variables

	Cases	Controls	log_2 [solCD44 (ng/ml)]					Protein (mg/ml)				
			Cases		Controls			Cases		Controls		
	N	N	Mean	SE	Mean	SE	P	Mean	SE	Mean	SE	P
All	150	150	**1.94**[a]	0.09	**1.28**[a]	0.07	<.0001	**0.94**[a]	0.05	**0.76**[a]	0.03	0.003
Site of enrollment												
JMH	80	71	**1.96**[a]	0.14	**1.32**[a]	0.11	<.0001	**0.95**[w]	0.07	**0.81**[w]	0.05	0.017
UM	70	79	**1.92**[b]	0.11	**1.26**[b]	0.09		**0.93**[a]	0.06	**0.73**[a]	0.04	
Age												
<60	84	85	**1.71**[a,c]	0.12	**1.16**[a,w]	0.08	<.0001	**0.88**[w]	0.06	**0.75**[w]	0.04	0.010
60 or more	66	65	**2.24**[b,c]	0.14	**1.45**[b,w]	0.12		**1.00**[a]	0.07	**0.78**[a]	0.05	
Gender												
Male	121	118	**2.01**[a]	0.10	**1.29**[a]	0.08	<.0001	**0.96**[a]	0.05	**0.80**[a]	0.04	0.006
Female	29	32	**1.68**	0.21	**1.28**	0.16		**0.86**[w]	0.10	**0.64**[w]	0.07	
Race/Ethnicity (n)	(149)	(148)										
White Non-Hispanic	53	29	**1.93**[a]	0.15	**1.31**[a]	0.14	<.001	**0.91**[a]	0.08	**0.68**[a]	0.07	0.017
White Hispanic	70	89	**1.91**[b]	0.13	**1.32**[b]	0.08		**0.91**	0.07	**0.81**	0.04	
Black	26	30	**2.06**[c]	0.26	**1.14**[c]	0.19		**1.08**[b]	0.12	**0.71**[b]	0.07	
SES												
Low	100	90	**1.92**[a]	0.12	**1.36**[a]	0.09	<.0001	**0.95**[w]	0.06	**0.81**[w]	0.04	0.009
High	50	60	**1.98**[b]	0.14	**1.17**[b]	0.10		**0.91**[a]	0.07	**0.69**[a]	0.04	
Smoking status												

Never	33	32	**1.72**[a]	0.20	**1.23**[a]	0.13	<.0001	**0.96**	0.14	**0.76**	0.06	0.027
Ever	117	118	**2.01**[b]	0.10	**1.30**[b]	0.08		**0.93**[a]	0.05	**0.76**[a]	0.04	
Never	33	32	**1.72**[a]	0.20	**1.23**[a]	0.13	<.0001	**0.94**	0.14	**0.76**	0.06	0.080
Former	37	59	**2.13**[b]	0.18	**1.31**[b]	0.10		**0.98**[a]	0.09	**0.78**[a]	0.06	
Current	80	59	**1.95**[c]	0.13	**1.29**[c]	0.12		**0.91**[w]	0.05	**0.75**[w]	0.05	
In current smokers (n)	(75)	(52)										
<20 pack-years	33	29	**1.86**[a]	0.18	**1.07**[a]	0.20	0.003	**0.96**	0.08	**0.71**	0.07	0.203
≥20 pack-years	42	23	**1.99**[w]	0.19	**1.51**[w]	0.14		**0.84**	0.08	**0.81**	0.09	
Alcohol past												
Non-drinker	35	40	**2.08**[a]	0.19	**1.38**[a]	0.13	<.0001	**1.00**[a]	0.09	**0.73**[a]	0.05	0.018
Drinker (Mild/Mod/Heavy)	115	110	**1.90**[b]	0.11	**1.25**[b]	0.08		**0.92**[b]	0.05	**0.78**[b]	0.04	
Alcohol current (n)	(148)	(148)										
Non-drinker	84	72	**1.87**[a]	0.12	**1.40**[a]	0.10	<.0001	**0.97**[w]	0.07	**0.82**[w]	0.05	0.010
Drinker (Mild/Mod/Heavy)	64	76	**2.04**[b]	0.15	**1.17**[b]	0.10		**0.91**[a]	0.06	**0.72**[a]	0.04	
Alcohol status												
Never	33	30	**2.08**[a]	0.20	**1.39**[a]	0.16	<.0001	**1.01**[a]	0.10	**0.75**[a]	0.07	0.019
Ever	117	120	**1.90**[b]	0.10	**1.26**[b]	0.08		**0.92**[b]	0.05	**0.77**[b]	0.04	
Teeth removed												
None/1 to 5	86	92	**1.82**[a,d]	0.11	**1.26**[a]	0.08	<.0001	**0.90**[a]	0.05	**0.74**[a]	0.04	0.020
≥6, but not all	36	39	**1.79**[b]	0.15	**1.25**[b]	0.14		**0.81**[b]	0.06	**0.76**	0.07	
All	24	15	**2.33**[c.d]	0.26	**1.43**[c]	0.20		**1.05**[b]	0.13	**0.84**	0.10	

Gargle (n)	(138)	(143)										
Poor/ Fair	38	12	**2.23**[a,c]	0.22	**0.88**[a]	0.36	<.0001	**1.13**[a,b]	0.11	**0.66**[a]	0.13	<0.001
Good	100	131	**1.82**[b,c]	0.10	**1.29**[b]	0.07		**0.85**[b]	0.05	**0.77**	0.03	
Cancer site												
Lip/OC	59		**2.12**	0.15			0.132	**0.98**	0.09			0.490
Oropharyngeal	91		**1.83**	0.11				**0.91**	0.05			
Stage												
Stage I/II	26		**1.78**	0.17			0.425	**0.90**	0.09			0.719
Stage III/IV	124		**1.98**	0.11				**0.94**	0.05			
T-stage												
T1-T2	63		**1.76**[w]	0.12			0.088	**0.89**	0.05			0.431
T3-T4	87		**2.07**[w]	0.13				**0.97**	0.07			
N-stage												
N0, Nx	51		**1.97**	0.14			0.848	**0.95**	0.08			0.850
N1-N3	99		**1.93**	0.12				**0.93**	0.06			
HPV (n)	(79)											
HPV+	31		**1.90**	0.23			0.760	**0.88**	0.09			0.997
HPV-	48		**1.99**	0.16				**0.88**	0.08			

SE: Standard error. P: p value from ANOVA global test of equality of all means. Same letter identify pairwise mean comparison within group or within a category of a key variable that was significant at the 5% level (letters a, b, c) or at the 10% level (letters w, y) by Fisher's least-significant-difference test. (n): total cases or total controls excluding missing data.

Table 3. Prediction models for oral cancer

Part A: Logistic Regression, all patients	Odds Ratio (95%CI)	*P*	AUC	Rescaled R^2
Univariate models (150 cases / 150 controls)				
log_2 solCD44	2.036 (1.552, 2.671)	<.0001	0.681	0.137
Protein	2.159 (1.288, 3.617)	0.003	0.590	0.042
Multivariable model [1] **(149 cases/148 controls)**				
log_2 solCD44	2.684 (1.797, 4.010)	<.0001	0.763	0.276
Protein	0.646 (0.301, 1.386)	0.262		
Part B: Logistic Regression stratified by HPV status				
HPV negative (48 cases / 150 controls)				
Univariate				
log_2 solCD44	2.311 (1.561, 3.422)	<.0001	0.689	0.146
Protein	1.838 (0.888, 3.807)	0.101	0.562	0.020
Multivariable model [2] (48 cases / 148 controls):				
log_2 solCD44	4.017 (2.124, 7.597)	<.0001	0.771	0.275
Protein	0.179 (0.052, 0.620)	0.006		
HPV positive (31 cases / 150 controls)				
Univariate				
log_2 solCD44	2.001 (1.291, 3.102)	0.002	0.667	0.096
Protein	1.882 (0.789, 4.492)	0.154	0.567	0.018
Multivariable model [3] (148 controls):				
log_2 solCD44	3.079 (1.486, 6.378)	0.003	0.773	0.221

Protein	0.384 (0.080, 1.833)	0.230

Part C: Logistic Regression Analysis of Risk Groups derived by Multivariate Recursive Partitioning

Univariate Model[4] of Risk Groups based on CD44 and protein levels

Risk Level (n = case + control)	**SolCD44**(ng/ml) (level description)	**Protein** (mg/ml)	**Odds Ratio (95%CI)**	**Prediction**	***P***	**AUC**	Rescaled R^2
Low (102 = 29 + 73)	<2.22 (low)	<1.23 (low-medium)	Reference	Control		0.722	0.227
Medium (116 = 54 + 62)	≥2.22 & <5.33 (medium)	≥0.558 (medium-high)	2.192 (1.247, 3.854)	Control	0.006		
High (5 = 4 + 1)	<2.22 (low)	≥1.23 (high)	10.069 (1.079, 93.93)	Case	0.043		
High (20 = 16 + 4)	≥2.22 & <5.33 (medium)	<0.558 (low)	10.069 (3.103, 32.672)	Case	0.0001		
High (57 = 47 + 10)	≥5.33 (high)	--	11.830 (5.279, 26.508)	Case	<.0001		

Multivariable Model[5] of Risk Groups based on CD44 and protein levels

Risk Level (n)	**SolCD44**	**Protein**	**Odds Ratio (95%CI)**	**Prediction**	***P***	**AUC**	Rescaled R^2
Low (102)	<2.22 (low)	<1.23 (low-medium)	Reference	Control		0.790	0.325
Medium (116)	≥2.22 & <5.33 (medium)	≥0.558 (medium-high)	2.755 (1.483, 5.117)	Control	0.001		
High (5)	<2.22 (low)	≥1.23 (high)	5.905 (0.591, 59.053)	Case	0.131		
High (20)	≥2.22 & <5.33 (medium)	<0.558 (low)	11.860 (3.312, 42.472)	Case	<.0001		
High (57)	≥5.33 (high)	--	14.489 (5.973, 35.145)	Case	<.0001		
SES High vs. low			0.577 (0.304, 1.094)		0.092		
White Non-Hispanic vs Black at age <60			7.885 (2.372, 26.206)				
White Hispanic vs Black at age <60			1.767 (0.636, 4.907)				
White Non-Hispanic vs Black at age ≥60			0.799 (0.216, 2.956)				

White Hispanic vs Black at age ≥60	0.382 (0.124, 1.175)
Age ≥60 vs <60 in Black	3.099 (0.838, 11.457)
Age ≥60 vs <60 in White Non-Hispanic	0.314 (0.111, 0.892)
Age ≥60 vs <60 in White Hispanic	0.669 (0.324, 1.383)
Alcohol Ever vs Never in Male	1.615 (0.713, 3.660)
Alcohol Ever vs Never in Female	0.202 (0.056, 0.726)
Male vs Female in alcohol=Never	0.216 (0.062, 0.757)
Male vs Female in alcohol=Ever	1.723 (0.695, 4.273)

AUC: area under the ROC curve. Rescaled R^2: coefficient of determination measured the dispersion explained by model. Odds ratios: 1-unit increase for continuous variables log2 CD44, protein, and age, unless specified categories; race/ethnicity (WNH and Black vs. WH), gender (Male v. Female), smoking and alcohol (Ever v. Never), and SES (high vs. low).

[1] Adjusted for age (p=0.132), race/ethnicity (p=0.004), age×race/ethnicity (p=0.006), gender (p=0.030), alcohol (p=0.032), gender×alcohol (p=0.020), smoking (p=0.527), and SES (p=0.042). Model "markers + covariates" (AUC=0.763) provided significantly better prediction than the reduced model excluding both markers (AUC=0.686) and only including potential risk factors (p=0.003), indicating that the markers aid prediction over and above prediction provided by knowledge of risk factors.

[2] Adjusted for age (p=0.020), gender (p=0.009), age×gender (p=0.008), race/ethnicity (p=0.740), alcohol (p=0.183), smoking (p=0.487), and SES (p=0.047).

[3] Adjusted for age (p=0.052), gender (p=0.104), age ×gender (p=0.096), race/ethnicity (p=0.298), alcohol (p=0.537), smoking (p=0.131), and SES (p=0.070).

[4] AUC=0.722 for risk group model (based on CD44 and protein) is significantly different from AUC =0.681 for univariate model $\log_2$ solCD44 (p=0.025).

[5] Logistic regression model included CD44-protein risk groups (5 categories, p<0.0001), age (≥60 vs. <60, p=0.090), gender (p=0.017), race/ethnicity (p=0.001), alcohol (p=0.014), SES (p=0.092), and interaction age×race/ethnicity (p=0.029) and gender×alcohol (p=0.007). Smoking (ever vs. never, p=0.700) and teeth removed (6 or more or all vs. 5 or less, p=0.485) were tested for inclusion into model (AUC=0.791); they were removed since their inclusion did not improve model fit.

Figure 1. 361 patients for the 2012 hospital-based case-control study were recruited from clinics at UM and JMH between 2007 and 2012. 19 patients were excluded from the study, 13 of those excluded patients did not meet inclusion criteria and another 6 were excluded from the study due to other reasons (patient withdrew consent or withdrew by PI discretion). 21 control patients were excluded from the main analysis for various reasons (potential for pre-malignant, pre-malignant, other cancers at time of enrollment). 17 cancer patients were excluded from main analysis for various reasons (unknown primary, other cancer, second primary at time of collection, study questionnaire was not available, tumor removed by biopsy before sample collection). 150 cancer cases and 150 controls met all inclusion criteria and were analyzed.

Figure 2. The differences in CD44 (**A**) and protein (**B**) levels over 1 year follow a normal distribution. Linear regression analysis shows that the trend towards decreasing levels over one year is significant for both CD44 (**C**) and protein (**D**). Kaplan-Meier Curves demonstrating significant differences in PFS (**E**) and OS (**F**) based on CD44 and protein level cutpoints.

Figure 1.

CONSORT 2010 Flow Diagram For Case-Control Study

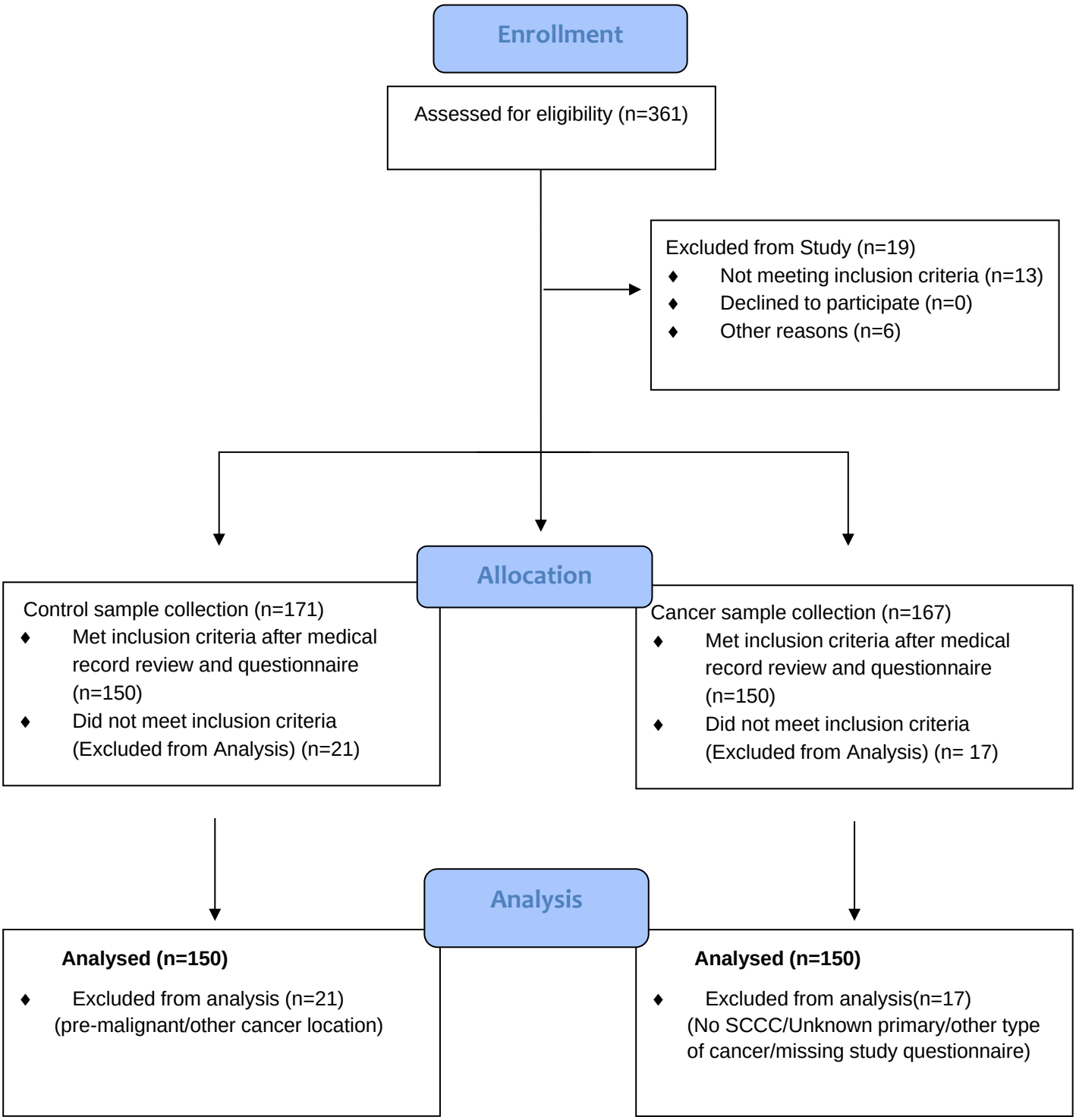
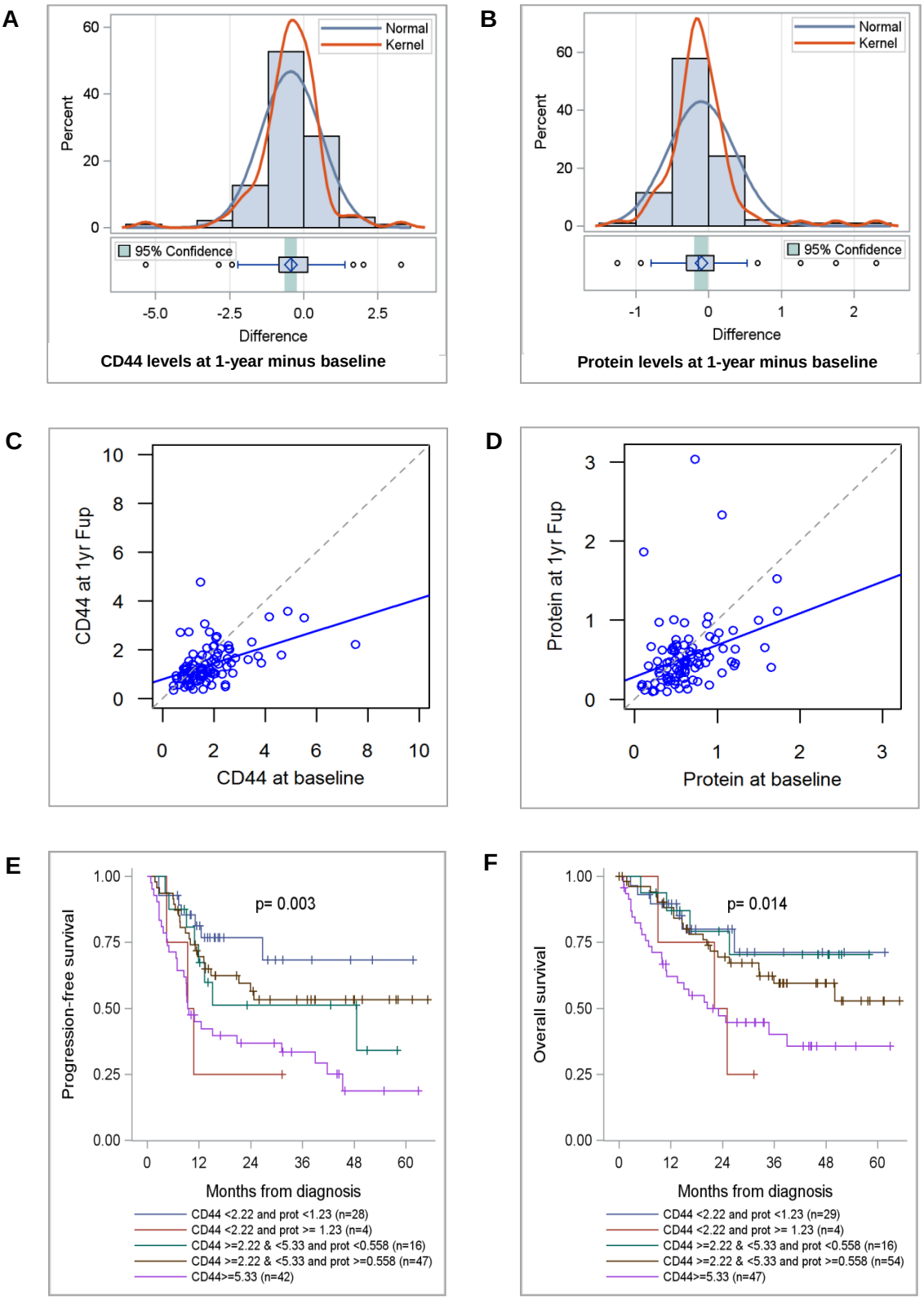


Figure 2.


A
Normal
Kernel
Percent
95% Confidence
Difference
CD44 levels at 1-year minus baseline
B
Protein levels at 1-year minus baseline
C
CD44 at 1yr Fup
CD44 at baseline
D
Protein at 1yr Fup
Protein at baseline
E
p= 0.003
Progression-free survival
Months from diagnosis
CD44 <2.22 and prot <1.23 (n=28)
CD44 <2.22 and prot >= 1.23 (n=4)
CD44 >=2.22 & <5.33 and prot <0.558 (n=16)
CD44 >=2.22 & <5.33 and prot >=0.558 (n=47)
CD44>=5.33 (n=42)
F
p= 0.014
Overall survival
CD44 <2.22 and prot <1.23 (n=29)
CD44 <2.22 and prot >= 1.23 (n=4)
CD44 >=2.22 & <5.33 and prot <0.558 (n=16)
CD44 >=2.22 & <5.33 and prot >=0.558 (n=54)
CD44>=5.33 (n=47)